TD3-based Adaptive Economic Dispatch Optimization Strategy for Multi-energy Microgrid

Jiakai Gong[1], Nuo Yu[1], Fen Han[1], Bin Tang[1], Haolong Wu[1], Yuan Ge[1]

***Abstract*—The multi-energy microgrid (MEMG) improves the overall economy of the system by coupling scheduling among multiple energy sources. However, in the case of renewable energy power generation and load demand fluctuations, traditional methods are difficult to apply to energy dynamic management and control under the changing situation of multi-energy microgrid systems, which poses a huge challenge to the multi-energy coupling optimal operation of MEMG. In this paper, a multi-energy allocation model based on deep reinforcement learning (DRL) is established to optimize the multi-energy coupling scheduling, which can automatically adapt to changes in the environment. In order to make the optimal scheduling strategy effectively reduce the cost, a multi-energy scheduling strategy based on the twin delayed deep deterministic policy gradient (TD3) algorithm is proposed. The experimental results display that our proposed strategy can reduce the cost by 21.45% and 14.71% compared with particle swarm optimization algorithm in summer and winter.**

***Index Terms*—Multi-Energy Microgrid (MEMG), Energy Dispatch, Deep Reinforcement Learning (DRL)**

I. INTRODUCTION

MULTI energy microgrids (MEMGs) are the emerging paradigms of large international energy enterprises, which is an organic conformity of distributed energy resources (DERs), energy coupling technologies, energy storage systems (ESSs), and loads demand on the distribution network level. The recent multiple energy research on energy coupling technologies carries the distributed generator (DG) equipment, and combined cooling, heating and power (CCHP) plants to provide power and thermal energy [1]. MEMGs have been proven to be an effective way to achieve economic operation and stable energy supply by improving the utilization rate of renewable energy (RESs) and coordinating multiple energy sources [2]. Therefore, one primary focus of academia is to explore the practical real-time MEMG energy scheduling problem under this circumstance [3].

Extensive research work has been conducted on optimal multi-energy coordination. Researchers mainly focus on the operation optimization of a single CCHP plant. These studies usually focus on two different dispatching strategies for power load and cooling/heating load. Literature [4] proposes a conversion strategy between electricity and cooling/heating loads, to comprehensively improve plant energy efficiency, reduce operating costs, and reduce carbon dioxide emissions. In addition, in order to effectively reduce the excess electricity or heat energy generated by CCHP plants, the researchers also compared the hybrid strategy of electricity and cold/heat loads with the strategy of single adherence to electricity or cold/heat loads [5]. In [6] and [7], the researchers designed the operation strategy of the CCHP plant according to different operation conditions and evaluation criteria. Although the operation optimization of a single CCHP has been extensively studied, the coordinated operation strategy of comprehensively considering CCHP power plants and other distributed generation resources in microgrid systems to supply multiple energy sources has not been fully studied. In order to optimize the complex multi-energy cooperative scheduling strategy, more flexible algorithms should be considered for the coordination of distributed generation equipment and CCHP supply.

In previous studies, most people have proposed many optimization algorithms for MEMG, which can be classified into two main techniques based on the approaches adopted: constrained optimization and heuristic algorithms. For the constrained optimization algorithms, reference [8] proposed the robust energy coordination system and applies mixed-integer linear programming (MILP) to minimize the operation cost for a novel hybrid AC/DC multi-energy ship (MES) MG with flexible thermal loads and voyage. An alternating direction method of multipliers (ADMM) algorithm is applied in [9], which is based on the average of the shared energy residual over all Multi-energy complementary MGs. Those methods are useful for a lot of complex tasks considering multiple factors and constraints. However, the MILP method assumes linear relationships among factors [10], while ADMM assumes that the problems are regularised and convex [11], which is unrealistic in many cases.

To address a large computational burden and constraints of constrained optimization algorithms for practical problems, most researchers study heuristic algorithms for optimizing MEMG scheduling problems. For example, An improved genetic algorithm is proposed to realize the economic and low-carbon operation of the system [12] - [13]. Reference [14] uses particle swarm optimization (PSO) approach to obtain a strategy that reduces the release of polluting gases and improves the overall energy expense. Reference [15] proposed an evolutionary algorithm (EA) to determine the scheduling model of a multi-energy hub. Even though these methods have been proven to be effective, several issues still hinder them from wider applications. The design of heuristic algorithms often depends on the designer's experience, and the lack of theoretical guidance may lead to the limitation of algorithm design. In addition, the large number of energy scheduling cooperations of MEMG may result in extremely high computational complexity leading to falling into a locally optimal solution.

In recent years, under the widespread application of artificial

Jiakai Gong and Nuo Yu are equal contribution to the paper. Nuo Yu is corresponding author (e-mail: yunuo@ahpu.edu.cn). 1. works with school of Electrical Engineering, Anhui Polytechnic University, Wuhu, China.

Manuscript received March 19, 2024, Revised April 23, 2024, Accepted April 30, 2024

intelligence technology [16], reinforcement learning (RL) has provided a new idea for solving the problems mentioned about MG energy scheduling in [17] - [18]. RL is a data-driven method. It has better adaptive learning ability and non-convex non-linear problem optimization decision-making ability. For example, the deep Q learning network (DQN) was used to solve the real-time dispatch strategy of the MG [19]. To solve the continuous action space optimization problem, the uncertain economic dispatch of the coupling energy storage based on the deep deterministic policy gradients (DDPG) approach is proposed [20]. However, the above algorithms are not only sensitive to the parameters and it is a daunting task to adjust the parameters, but also have the problem of overestimated Q values. Therefore, how to make reinforcement learning effectively evaluate the value function of MEMG to obtain the optimization strategy with the lowest operating cost needs to be studied at present.

The multi-energy cooperative scheduling problem contains many decision variables and complex constraints. In addition, renewable energy, load, electricity and natural gas prices need to be considered for system optimization operating costs, which can lead to computational challenges for existing commercial solvers and heuristic algorithms when dealing with such problems. However, the current reinforcement learning solution to microgrid scheduling will overestimate the value function, resulting in the inability to learn the optimal strategy. To fill the existing research gaps identified above, this paper presents a MEMG scheduling strategy based on the twin delayed deep deterministic policy gradient algorithm [21] to minimize system operation cost.

The main contributions of this paper are summarized as follows:

1) The multi-energy scheduling optimization model with minimum operating costs is transformed into an adaptive reinforcement learning Markov decision process (MDP) model, which overcomes the complex problem of establishing an accurate optimization model for multi-energy coupling with fluctuations in renewable energy and multiple types of loads.

2) To solve the problem of overestimation of the value function of the microgrid scheduling strategy in reinforcement learning, the MEMG scheduling strategy based on a twin delayed deep deterministic policy gradient (TD3) algorithm is proposed. It can effectively learn optimal scheduling strategies through self-adaptive environments.

3) The simulation results show that our method can obtain better solutions of convergence and stability, and the proposed MEMG dispatch strategy can reduce the cost by 21.45% and 14.71% compared with particle swarm optimization algorithm in summer and winter.

The rest of the paper is concluded as follows. The definition of MEMG scheduling decision-making is presented in Section II. The energy management of MEMG is introduced in Section III. The MDP optimization model of MEMG is introduced in Section IV. Section V describes the TD3 approach. Simulation results for a real case study are presented in Section VI. The conclusion is drawn in Section VII.

II. MULTI-ENERGY SYSTEM MODEL

A. System Framework

The structure of MEMG is shown in Fig. 1. The renewable energy contains WT and PV, where the WT and PV are non-dispatchable generators with high intermittence and fluctuation [22]. Air conditioners (AC) and electric coolers (EC) are the Heat/cold energy conversion equipment. CCHP and fuel cell (FC) are dispatchable generators. The load demand consists of heat, cold, and electric loads. ESS includes battery energy storage (BES), heat energy storage (TES), and cool energy storage (CES). The MEMG is connected to the main grid for electrical energy exchange and to the gas company to provide heat/cooling load demand through a gas boiler (GB) that converts natural gas into heating/cooling power.

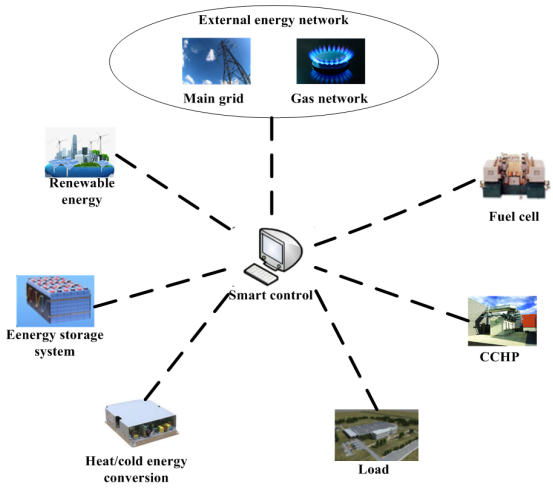


Fig. 1: MEMG system framework.

B. CCHP System

The CCHP equipment contains an MT, heat exchanger, and absorption chiller. The waste thermal/cold can be recovered through an absorption chiller/heat exchanger and generate thermal/cold energy. The gas consumption of MT can be formulated as follows [23]:

$$V_{\mathrm{MT}}^{t}=\frac{P_{\mathrm{MT}}^{t}}{\eta_{\mathrm{MT}} L_{\mathrm{NG}}} \Delta t \tag{1}$$

where t is the time of system dispatch; V_{MT}^{t} is the gas consumption volume of MT at time t; P_{MT}^{t} and η_{MT} are the power outputs and coefficient of MT; L_{NG} is the low calorific value; Δt is the scheduling time unit.

The generated thermal/cold of MT can be described as follows [23]:

$$\begin{cases} Q_{\mathrm{MT}}^{t}=\frac{P_{\mathrm{MT}}^{t}\left(1-\eta_{\mathrm{MT}}-\eta_{\mathrm{L}}\right)}{\eta_{\mathrm{MT}}} \\ Q_{\mathrm{MTC}}^{t}=\eta_{\mathrm{rec}} \eta_{\mathrm{MTC}} Q_{\mathrm{MT}}^{t} \\ Q_{\mathrm{MTH}}^{t}=\eta_{\mathrm{rec}} \eta_{\mathrm{MTH}} Q_{\mathrm{MT}}^{t} \end{cases} \tag{2}$$

where Q_{MT}^{t} is the output power of heat wasted of MT; η_{L} is the thermal loss rate; Q_{MTC}^{t} and Q_{MTH}^{t} are the output power of cold and thermal energy, respectively; η_{MTC} and η_{MTH} are the cold/heat output efficiency. η_{rec} is the recovery ratio of cold/heat energy.

C. Fuel cell

FC can provide electric power through gas consumption. The natural gas consumption volume of FC can be formulated as follows:

$$V_{\mathrm{FC}}^{t}=\frac{P_{\mathrm{FC}}^{t}}{\eta_{\mathrm{FC}}L_{\mathrm{NG}}}\Delta t \tag{3}$$

where V_{FC}^{t} is the gas volume consumed; P_{FC}^{t} is power outputs of FC; η_{FC}^{t} is the electric conversion efficiency of FC.

D. Gas boiler

The energy conversion of GB generates thermal/cooling power by consuming natural gas. Its heat power output model is as follows:

$$Q_{\mathrm{GB}}=\eta_{\mathrm{GB}}V_{\mathrm{GB}}L_{\mathrm{NG}} \tag{4}$$

where Q_{GB} and V_{GB} are heat production and gas consumption of the GB, respectively; η_{GB} represents the efficiency of heat production of the GB.

E. Electric chiller and air conditioner

EC and AC are energy conversion equipment that can convert electric energy into cold and thermal energy respectively. The output power of cold and heat can be modeled as follows:

$$Q_{\mathrm{EC}}^{t}=\eta_{\mathrm{EC}}P_{\mathrm{EC}}^{t} \tag{5}$$

$$Q_{\mathrm{AC}}^{t}=\eta_{\mathrm{AC}}P_{\mathrm{AC}}^{t} \tag{6}$$

where P_{EC}^{t} and P_{AC}^{t} represent the output power of EC and AC; η_{EC} and η_{AC} are the energy conversion efficiency of cold and thermal; Q_{EC}^{t} and Q_{AC}^{t} are the cold and thermal output power of EC and AC, respectively.

F. Energy storage system

In our work, we consider the ESSs to have three types: BES, TES, and CES. The ESS mathematical model is mainly related to the SOC of ESS, which can be formulated as follows [24]:

$$\begin{aligned}SOC_{\mathrm{i}}^{t}=(1-\tau)SOC_{\mathrm{i}}^{t-1}&+\frac{P_{\mathrm{i,ch}}^{t}\eta_{\mathrm{ch}}}{E_{\mathrm{i}}}\Delta t\\&-\frac{P_{\mathrm{i,dis}}^{t}/\eta_{\mathrm{dis}}}{E_{\mathrm{i}}}\Delta t\end{aligned} \tag{7}$$

where i is the indication of ESSs; SOC_{i}^{t} is the SOC of i th ESS at time t, respectively; τ is the decay rate of SOC; $P_{\mathrm{i,ch}}^{t}/P_{\mathrm{i,dis}}^{t}$ are the charging/discharging power of i th ESS, respectively; E_{i} is the rated capacity of i th ESS; η_{ch} and η_{dis} are the charging and discharging rates of ESS; Δt represents the time step.

III. OPTIMAL OPERATION MODEL

The specific model of the MEMG dispatch process is introduced as follows.

A. Object Function

MEMG energy dispatch aims to minimize the system operation cost, which consists of fuel cost, maintenance cost, power exchange cost, and start-stop cost in four parts. The objective function is as follows

$$F_{\mathrm{G}}=\sum_{t=1}^{T}C_{\mathrm{FU}}^{t}+C_{\mathrm{ME}}^{t}+C_{\mathrm{EX}}^{t}+C_{\mathrm{ST}}^{t} \tag{8}$$

$$C_{\mathrm{FU}}^{t}=P_{\mathrm{gas}}^{t}(V_{\mathrm{MT}}^{t}+V_{\mathrm{FC}}^{t}+V_{\mathrm{GB}}^{t}) \tag{9}$$

$$C_{\mathrm{ME}}^{t}=\begin{bmatrix}K_{\mathrm{WT}}^{\mathrm{me}}P_{\mathrm{WT}}^{t}+K_{\mathrm{PV}}^{\mathrm{me}}P_{\mathrm{PV}}^{t}+K_{\mathrm{MT}}^{\mathrm{me}}P_{\mathrm{MT}}^{t}\\+K_{\mathrm{FC}}^{\mathrm{me}}P_{\mathrm{FC}}^{t}+K_{\mathrm{EC}}^{\mathrm{me}}P_{\mathrm{EC}}^{t}+K_{\mathrm{AC}}^{\mathrm{me}}P_{\mathrm{AC}}^{t}\\+K_{\mathrm{BES}}^{\mathrm{me}}(P_{\mathrm{BES,ch}}^{t}+P_{\mathrm{BES,dis}}^{t})\\+K_{\mathrm{TES}}^{\mathrm{me}}(P_{\mathrm{TES,ch}}^{t}+P_{\mathrm{TES,dis}}^{t})\\+K_{\mathrm{CES}}^{\mathrm{me}}(P_{\mathrm{CES,ch}}^{t}+P_{\mathrm{CES,dis}}^{t})\end{bmatrix}\Delta t \tag{10}$$

$$C_{\mathrm{EX}}^{t}=(K_{\mathrm{pur}}^{t}P_{\mathrm{grid,pur}}^{t}+K_{\mathrm{sell}}^{t}P_{\mathrm{grid,sell}}^{t})\Delta t \tag{11}$$

$$C_{\mathrm{ST}}^{t}=\sum_{j=1}^{M}\max\{0,U_{\mathrm{GD},j}^{t}-U_{\mathrm{DG},j}^{t-1}\}C_{\mathrm{ST},j}\Delta t \tag{12}$$

where F_{G} is the total operating cost of MEMG system; T is the dispatch times; where F_{G} and T are the total operation cost and the scheduling periods; C_{FU}^{t}, C_{ME}^{t}, C_{EX}^{t}, and C_{ST}^{t} are the cost of fuel, maintenance, energy exchange, start-stop; P_{gas}^{t} is the natural gas price; V_{GB}^{t} is the volume of gas consumed; $K_{\mathrm{WT}}^{\mathrm{me}}$, $K_{\mathrm{PV}}^{\mathrm{me}}$, $K_{\mathrm{BES}}^{\mathrm{me}}$, $K_{\mathrm{TES}}^{\mathrm{me}}$, $K_{\mathrm{CES}}^{\mathrm{me}}$, $K_{\mathrm{MT}}^{\mathrm{me}}$, $K_{\mathrm{FC}}^{\mathrm{me}}$, $K_{\mathrm{EC}}^{\mathrm{me}}$ and $K_{\mathrm{AC}}^{\mathrm{me}}$ are the unit maintenance cost of WT, PV, BES, TES, CES, MT, FC, EC and AC; K_{pur}^{t} and K_{sell}^{t} are the unit price for purchase and sell; $P_{\mathrm{grid,pur}}^{t}$ and $P_{\mathrm{grid,sell}}^{t}$ are the purchase and sell power; M represents the number of the dispatchable equipment; j represents the j-th dispatchable units; $U_{\mathrm{DG,j}}$ is the binary values with values of 0 and 1; $C_{\mathrm{ST,j}}$ is the start-stop cost of j-th equipment.

B. Constraints

The system possibly becomes unsafe and unstable when MEMG is operating. Therefore system model constraints are considered necessary.

1) Power Balance Constraints: The equipment output power should satisfy the load demands. The power balance constraints of system at time step t can be expressed as

$$\begin{aligned}&P_{\mathrm{MT}}^{t}+P_{\mathrm{FC}}^{t}+P_{\mathrm{WT}}^{t}+P_{\mathrm{PV}}^{t}+P_{\mathrm{grid,pur}}^{t}-P_{\mathrm{grid,sell}}^{t}+\\&P_{\mathrm{BES,dis}}^{t}-P_{\mathrm{BES,ch}}^{t}=P_{\mathrm{load}}^{t}+P_{\mathrm{EC}}^{t}+P_{\mathrm{AC}}^{t}\end{aligned} \tag{13}$$

$$Q_{\mathrm{MTC}}^{t}+Q_{\mathrm{EC}}^{t}+P_{\mathrm{CES,ch}}^{t}-P_{\mathrm{CES,dis}}^{t}+Q_{\mathrm{GB}}^{t}=Q_{\mathrm{cool}}^{t} \tag{14}$$

$$Q_{\mathrm{MTH}}^{t}+Q_{\mathrm{AC}}^{t}+P_{\mathrm{TES,dis}}^{t}-P_{\mathrm{TES,ch}}^{t}+Q_{\mathrm{GB}}^{t}=Q_{\mathrm{heat}}^{t} \tag{15}$$

where P_{load}^{t} is the load demands of the system at time t; P_{WT}^{t} and P_{PV}^{t} are the output power of WT and PV, respectively; P_{load}^{t}, Q_{cool}^{t}, and Q_{heat}^{t} are the loads demand of electricity, cold, and thermal, respectively.

2) Power constraints of equipment: The electric/heat/cod power of device operation should have upper and lower limits to prevent device damage.

$$P_{\mathrm{i}}^{\min} \leq P_{\mathrm{i}}^{t} \leq P_{\mathrm{i}}^{\max} \tag{16}$$
$$Q_{\mathrm{i}}^{\min} \leq Q_{\mathrm{i}}^{t} \leq Q_{\mathrm{i}}^{\max} \tag{17}$$

where P_{i}^{t} is the output power of the device i that provide electric energy; Q_{i}^{t} is the heat/cold output power of the device i; $P_{\mathrm{i}}^{\min}$ and $P_{\mathrm{i}}^{\max}$ are the minimum and maximum power of device i that supply electric energy; $Q_{\mathrm{i}}^{\min}$ and $Q_{\mathrm{i}}^{\max}$ are the minimum and maximum power of device i that supply heat/cold energy;

3) Energy Storage system Constraints: The level of SOC may influence the service life of energy storage. Therefore, in addition to power limitation, we should also consider whether SOC is within the reasonable use range.

$$SOC_{\mathrm{i}}^{\min} \leq SOC_{\mathrm{i}}^{t} \leq SOC_{\mathrm{i}}^{\max} \tag{18}$$
$$P_{\mathrm{i,ch}}^{\min} \leq P_{\mathrm{i,ch}}^{t} \leq P_{\mathrm{i,ch}}^{\max} \tag{19}$$
$$P_{\mathrm{i,dis}}^{\min} \leq P_{\mathrm{i,dis}}^{t} \leq P_{\mathrm{i,dis}}^{\max} \tag{20}$$

where $SOC_{\mathrm{i}}^{\min}$ and $SOC_{\mathrm{i}}^{\max}$ are the minimum and maximum SOC of ESS i; $P_{\mathrm{i,ch}}^{\min}$ and $P_{\mathrm{i,ch}}^{\max}$ are minimum and maximum charging power of ESS i; $P_{\mathrm{i,dis}}^{\min}$ and $P_{\mathrm{i,dis}}^{\max}$ are minimum and maximum discharging power of ESS i.

IV. REINFORCEMENT LEARNING FRAMEWORK FOR MEMG DISPATCH PROBLEM

This decision-making process is constructed as a Markov decision process (MDP) [25]. The MDP consists of (S, A, T, R), whose design is as follows.

A. State space

The states of the MEMG energy dispatch include the output power of WT and PV, the state of charge of BES, TES, and CES, the load of electric, heating, and cooling, and the price of purchase and sell power in utility. Therefore, the state space can be defined as

$$s_t = [P_{\mathrm{WT}}^{t}, P_{\mathrm{PV}}^{t}, SOC_{\mathrm{BES}}^{t}, SOC_{\mathrm{TES}}^{t}, SOC_{\mathrm{CES}}^{t}, P_{\mathrm{load}}^{t}, Q_{\mathrm{heat}}^{t}, Q_{\mathrm{cool}}^{t}, P_{\mathrm{grid,pur}}^{t}, P_{\mathrm{grid,sell}}^{t}] \tag{21}$$

B. Action Space

The action space is the decisions made by the agent. The action of the MEMG energy dispatch decision contains the power of dispatchable DGs, conversion energy units, and ESS output power. The action space can be defined as

$$a_t = [P_{\mathrm{MT}}^{t}, P_{\mathrm{FC}}^{t}, P_{\mathrm{EC}}^{t}, P_{\mathrm{AC}}^{t}, P_{\mathrm{BES}}^{t}, P_{\mathrm{TES}}^{t}, P_{\mathrm{CES}}^{t}] \tag{22}$$

C. State Transition Probability

The state transition probability is that the agent will move from the current state s_t to the new state s_{t+1} after performing the selected action a_t.

D. Reward Function

The objective function of the MEMG energy scheduling problem is to decrease the operation cost. The minimization of the MEMG dispatch operation cost is transformed into the maximization of the reward function. The reward function at the time step t can be defined as follows:

$$r_t = -(C_{\mathrm{FU}}^{t} + C_{\mathrm{ME}}^{t} + C_{\mathrm{EX}}^{t} + C_{\mathrm{ST}}^{t}) \tag{23}$$

V. REINFORCEMENT LEARNING ALGORITHM FOR MEMG DISPATCH PROBLEM

A. Reinforcement Learning Preliminary

RL considers agents constantly learning from the environment to obtain reward-maximizing behavior or a given goal, whose process is shown in Fig. 2. For a given state $s_t \in S$,

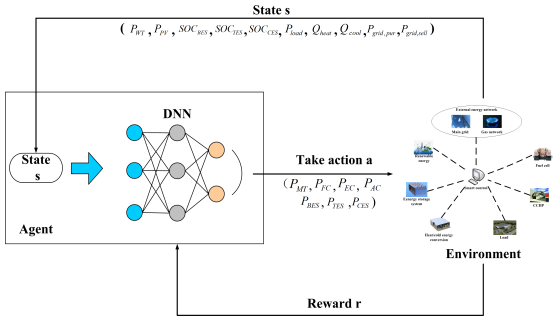


Fig. 2: Interaction process between agent and environment.

the agent selects action a_t in the action space A according to a policy π. The policy can be mapped from s_t to a_t. Then the environment will return a reward value r_t and the next time state s_{t+1} after executing the action a_t. The agent and environment will continue to learn through interaction until the time series ends. The cumulative reward of discounted can be formulated as follows:

$$R_t = \sum_{i=t}^{T} \gamma^{i-t} r(s_i, a_i), \tag{24}$$

where t indicates the time step, T is the length of an episode and γ is a discount factor determining the priority of short-term rewards, with a range of (0,1].

The optimal strategy of the goal can be solved by calculating the maximum expected cumulative discount reward. The action value function represents the expected cumulative discount reward at the state s_t after adopting the strategy π, and can be described as follows:

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{\pi}[R_t | s_t, a_t]. \tag{25}$$

B. TD3 Algorithm

The TD3 algorithm adopts the following three techniques to solve the problem of excessive estimation bias, the framework of our proposed algorithm is depicted in Fig. 3. First, two sets of critic networks are employed to evaluate the Q value in the TD3 algorithm. As shown in (26), the smaller one is selected to update the target Q value, to alleviate the overestimation of Q value. The loss function is defined as the square of the

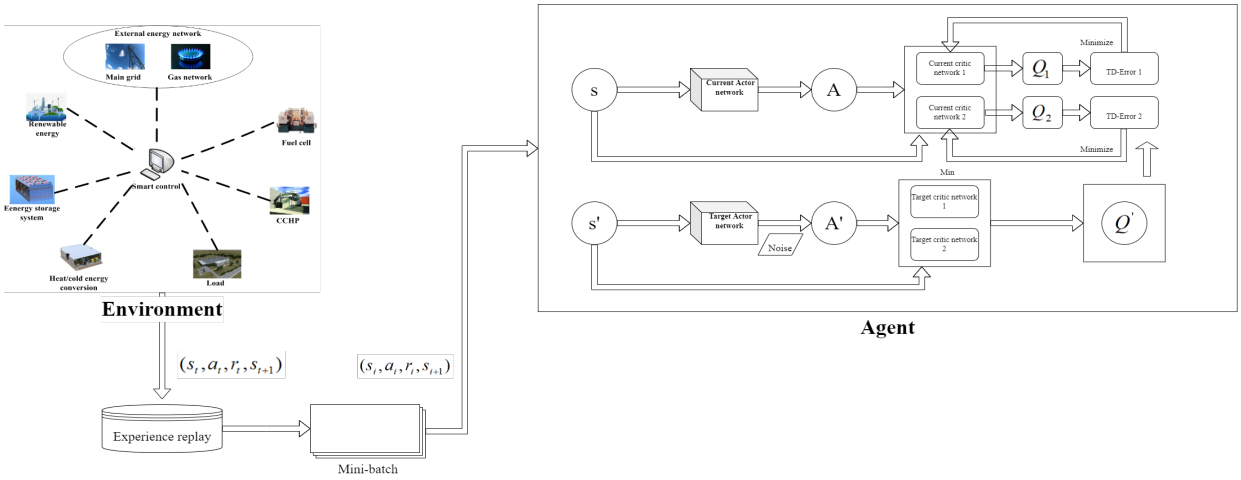


Fig. 3: TD3 algorithm framework.

difference between the selected target Q value and the neural network output Q value, which is defined as (27).

$$y\left(r, s'\right) = r + \gamma \min_{i=1,2} Q_{\theta_i'}\left(s', a'\right) \tag{26}$$

$$L\left(\theta_i\right) = E\left[\left(Q_{\theta_i}(s,a) - y\left(r, s'\right)\right)^2\right] \tag{27}$$

where s', a' and r are the next state, next action and reward, γ is discount factor. θ and θ' represent current and target actor network parameters respectively.

Algorithm 1 TD3.

Initialize replay buffer D with capacity N, iterative times M, discount factor γ, sample batches size S, critic networks Q_{θ_1}, Q_{θ_2} and actor network π_ϕ with random weights θ_1, θ_2, ϕ
Initialize target networks $\theta_1' \leftarrow \theta_1$, $\theta_2' \leftarrow \theta_2$, $\phi' \leftarrow \phi$
for $episode$ = 1, ..., M **do**
  Initial state s_0
  for t = 1 to T **do**
    Select action with exploration noise:
    $a \sim \pi_\theta(s) + \sigma, \sigma \in \mathcal{N}(0, \widetilde{\omega})$
    Observe reward r and new state s'
    Store transition tuple (s, a, r, s') in D
    Sample minibatch of transitions (s, a, r, s') from D
    $a' \sim \pi_\theta'(s') + \sigma, \sigma \in clip(\mathcal{N}(0, \widetilde{\omega}), -c, c)$
    $y = r + \gamma \min_{i=1,2} Q_{\theta_i'}(s', a')$
    Update critics:
    $\theta_i \leftarrow \arg\min_{\theta_i} \ N^{-1} \sum \left(y - Q_\theta\left(s, a\right)\right)^2$
    if t mod d **then**
      Update ϕ by the deterministic policy gradient:
      $\nabla_\phi J\left(\phi\right) = N^{-1} \sum \nabla_a Q_{\theta_i}\left(s, a\right)|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi\left(s\right)$
      Update target networks:
      $\theta_i' \leftarrow \tau\theta_i + \left(1-\tau\right)\theta_i'$
      $\phi' \leftarrow \tau\phi + \left(1-\tau\right)\phi'$
    end if
  end for
end for

Second, the update frequency of the actor network is lower than that of the critic network, which will not be updated unless the value function changes significantly. The delayed policy update will result in a lower variance of the value estimate and thus form a stable policy. Then, the critic network is stable and has fewer errors before being used to update the actor network. The soft update is adopted in TD3's target network, which can be expressed as:

$$\begin{aligned} \theta_i' &\leftarrow \tau\theta_i + \left(1-\tau\right)\theta_i' \\ \phi' &\leftarrow \tau\phi + \left(1-\tau\right)\phi' \end{aligned} \tag{28}$$

where τ is the update coefficient. θ is the current critic network parameter. ϕ is the current actor network parameter. θ' and ϕ' are critic and actor target network parameters.

Third, learning objectives using deterministic policy are highly susceptible to inaccuracies caused by function approximation errors when updating the actor network and then increasing the variance of the objective. To mitigate this problem, the TD3 algorithm adds noise to the action and averages over a size of mini-batch N to smooth the estimates

$$\widetilde{a} = \pi_{\phi'}\left(s'\right) + \sigma, \quad \sigma \sim clip\left(\mathcal{N}\left(0, \widetilde{\omega}\right), -c, c\right) \tag{29}$$

where the added noise σ is limited to the range of $(-c, c)$ to ensure that the deviation of the processed action is controllable. The updated principle of the TD3 algorithm is

$$\theta_i \leftarrow \arg\min_{\theta_i} \ N^{-1} \sum \left(y - Q_\theta\left(s, a\right)\right)^2 \tag{30}$$

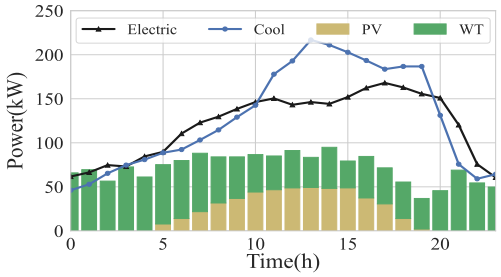
$$\nabla_\phi J\left(\phi\right) = N^{-1} \sum \nabla_a Q_{\theta_i}\left(s, a\right)|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi\left(s\right). \tag{31}$$

The exact procedure of the TD3 algorithm in this research is presented in the following Algorithm 1

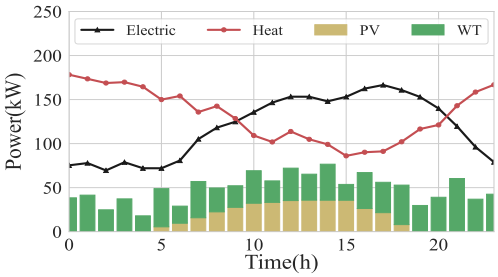
VI. CASE STUDY

A. Experimental Setting

We consider the heat load demand is only required in winter, and the cold load demand is only required in summer. The load demands, PV, and WT power outputs [26] are shown in Fig. 4. The scheduling period is set to one day, which is divided into 24-time steps. Table I introduces the equipment power limits



(a) summer



(b) winter

Fig. 4: Experiment data.

TABLE I: Maintenance Cost

Unit Type	Power out(kW)		Unit Maintenance Cost(¥/kW)
	MIN	MAX	
MT	15	65	0.025
FC	0	40	0.028
EC	0	30	0.026
AC	0	30	0.026
WT	0	65	0.026
PV	0	50	0.025
BES	-30	30	0.013
TES/CES	-20	20	0.013
GB	0	65	-
Grid	-60	60	-

TABLE II: Parameter of Energy Storage Device

Type	BES	TES/CES	Type	BES	TES/CES
τ	0.001	0.001	$\lambda_{\min}$	0.2	0.2
$\eta_{\mathrm{ch}}/\eta_{\mathrm{dis}}$	0.95/1.05	0.95/1.05	$\lambda_{\max}$	0.9	0.9
$P_{\mathrm{ch,max}}$	30	20	SOC^0_{ES}	0.2	0.2
$P_{\mathrm{dis,max}}$	30	20	E_{ES}	150	150

and maintenance costs for the MEMG system. Table II reveals the ESSs' equipment parameters. Table III shows the time-of-use price. The start-stop costs for devices of MT, FC, EC, and AC are assumed by 1.94, 2.21, 1.32, and 1.36, respectively in this work. Table IV exhibits the generation coefficient of energy conversion. The L_{NG} is set to 9.7 kW·h/m^3; the gas price set as 2.05 ¥/m^3.

B. Algorithm Performance Analysis

The performance of the three benchmark RL methods is compared with our used method in the system scheduling problem on typical days [25].

The comparative changes diagram of the cumulative reward function of four algorithms in the training process of 5000 is

TABLE III: Electricity Price

Type	Time Period	Purchase Price (¥/kWh)	Sell Price (¥/kWh)
Peak	7.00-10.00 18.00-21.00	0.98	0.50
Normal	10.00-18.00 21.00-23.00	0.49	0.20
Valley	23.00-07.00	0.17	0.00

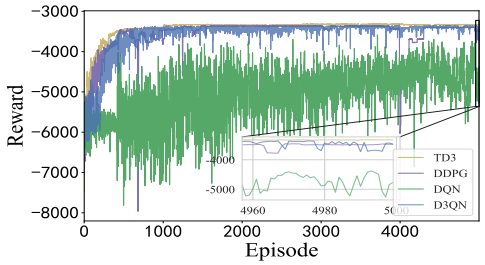
TABLE IV: Coefficient of Distributed Generation

Parameters	Value	Parameters	Value
η_{rec}	0.95	η_{EC}	0.95
η_{GB}	0.80	η_{AC}	0.95
η_{MTC}	0.90	η_{MT}	0.45
η_{MTH}	0.90	η_{FC}	0.85

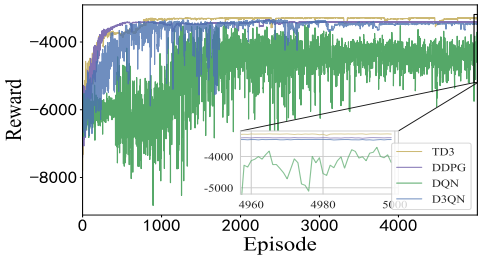
TABLE V: Algorithm Performance

Data	Algorithm	performance		
		Deviation	Average	Maximum
Summer	TD3	328.92	-3394.54	-3275.69
	DDPG	461.63	-3563.71	-3352.85
	DQN	661.62	-5068.12	-3585.66
	D3QN	472.00	-3567.71	-3336.07
Winter	TD3	361.50	-3431.41	-3278.87
	DDPG	386.39	-3513.42	-3341.26
	DQN	874.24	-4865.15	-3478.75
	D3QN	553.16	-3674.80	-3329.60

shown in Fig. 5. It can be seen that the reward obtained through the four methods has increased constantly at the beginning. This is because the agent interacts with the environment and can study the optimizing dispatch strategy. The reward obtained from our method is convergence after around 1100 and around 900 train episodes and stable at about 3285 and 3280 in summer and winter respectively. Table 5 shows several

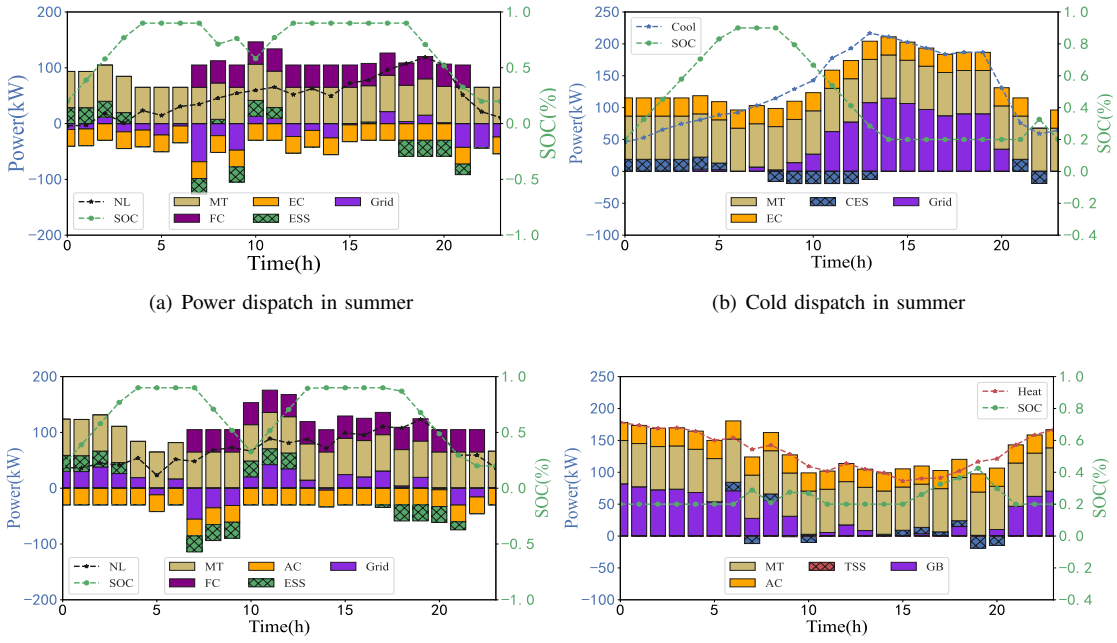


(a) summer



(b) winter

Fig. 5: Reward.



(a) Power dispatch in summer

(b) Cold dispatch in summer

(c) Power dispatch in winter

(d) Heat dispatch in winter

Fig. 6: Energy scheduling.

performance metric results of different algorithms on summer and winter days. The average and maximum reward values obtained through our proposed method are the largest, and the standard deviation is 328.92, which is smaller than other methods.

From the comparative results shown in Fig. 5 and Tabel V, it can be seen that the TD3 algorithm has the best convergence performance, and the system dispatch strategy obtained is more stable than others in the training process.

C. Dispatch Result Analysis

In order to verify the effectiveness of the proposed DRL-based scheduling model using the TD3 algorithm training strategy

The electric and cooling energy scheduling results on typical days are revealed in Fig. 6. In those figures, NL represents the net load that electrical load minus WT and PV outputs. Grid is the electric energy exchanged between the MEMG system and the utility.

Combined with the analysis in Fig. 6, Fig. 4, and Table 3. The energy dispatch strategy of TD3 in summer can make orderly outputs under the guidance of external electricity prices. Fig. 6(a) and 6(c) show the electric energy dispatch strategy of TD3. When the electrical price is at a valley period, which is lower than other periods, and the power costs generated by FC are relatively high, the MEMG system mainly purchases electricity to satisfy the electricity load demand. Due to the need to meet the cold or heat load requirements of the system, MT can provide cold and heat energy for the system at maximum power, reducing high-cost purchased energy. Besides, the BES does a lot of charging until the SOC is a maximum of 0.9 for subsequent use during peak hours. When the electricity price is relatively high, the BES discharges a part of the electricity. The generated power costs of FC and CCHP are lower than the purchase electricity price, and FC and CCHP start outputting maximum power. At the same time, the MEMG system sells the electricity to the main grid, which can reduce the electrical exchange cost with utility.

Fig. 6 (b)/6(d) shows the cold/heat energy dispatch strategy of TD3. The cooling power of the MEMG system is mainly supplied by CCHP, which reserves the waste cold/heat in the electricity network for the cold/heat network. Due to the output efficiency of EC/AC being higher than GB, the system firstly relies on the EC/AC to convert electricity from the electric network into cooling/heating outputs to satisfy the load demands. The GB outputs power when the energy supply of the system is insufficient. When the load is relatively low and the energy provided by CCHP and EC/AC is excessive, the storage system starts to charge a lot, then carries out a large number of discharges when the system energy shortage to avoid purchasing more natural gas. In addition, Table VI shows the operation costs of dispatch results, the TD3-based dispatch strategy can achieve minimum operation cost compared with others. The simulation results demonstrate proposed dispatch strategy can reduce the cost by 21.45% and 14.71% compared with PSO in summer and winter.

VII. CONCLUTION

For the MEMG system to minimize the operating cost, this paper proposes to transform the MEMG system scheduling optimization model into a reinforcement learning Markov decision process under self-adaptive environmental changes,

TABLE VI: Operation Costs of MEMG

Cost/revenue (¥)	Summer					Winter				
	TD3	DDPG	DQN	D3QN	PSO	TD3	DDPG	DQN	D3QN	PSO
Fuel cost	3208.61	3200.60	4350.97	3308.40	3939.25	3081.34	3162.09	3756.82	3210.34	3526.74
Maintenance cost	126.42	118.80	107.65	122.70	113.45	112.69	111.44	98.62	115.37	110.61
Power exchange	-48.70	152.55	312.43	39.34	125.46	84.68	121.92	337.14	133.91	205.16
Start-stop cost	2.21	4.42	12.80	6.63	8.84	2.21	2.21	10.59	2.21	4.42
Total costs	3288.55	3476.37	4783.85	3477.08	4387.00	3280.92	3397.66	4203.16	3461.83	3846.93

which avoids the difficulty of establishing accurate models under real-time changes in the environment, then proposes an optimal scheduling strategy for MEMG system based on the TD3 algorithm, which solves the problem that the traditional reinforcement learning evaluation Q value function of microgrid strategy is overestimated and leads to having a deviation for finding the optimal scheduling strategy, and realizes the self-learning optimal scheduling strategy with the lowest operating cost under real-time changes. Compared with the PSO method, the TD3-based dispatch strategy can reduce the cost by 21.45% and 14.71% in summer and winter, respectively.

In this paper, the MEMG self-learning optimization scheduling strategy based on reinforcement learning solves the problem of the lowest operating cost of the system, and the system learns the optimization strategy through interaction with the environment, and the internal working mechanism of the model and algorithm is opaque to the final decision maker. It makes it difficult to understand the system operating mechanism and cannot guarantee the safe operation of the system. The key to future work is how to realize the MEMG optimization strategy scheduling based on interpretable reinforcement learning.

ACKNOWLEDGMENTS

This work is supported by Anhui Provincial Natural Science Foundation (No. 2008085MF208), Open Research Fund of Anhui Key Laboratory of Detection Technology and Energy Saving Devices, Anhui Polytechnic University (No. JCKJ2021A07), National Natural Science Foundation of China (NSFC) (No. U21A20146), and Collaborative Innovation Project of Anhui Universities (No. GXXT-2020-070).

REFERENCES

[1] L. Chen, J. Wu, F. Wu, H. Tang, and Y. Xiong. Energy flow optimization method for multi-energy system oriented to combined cooling, heating and power. Energy, pp. 118–536, 2020.

[2] Y. Xu, L. Wu, S. L. Walker, J. Lian, A. Verma, and R. Zhang. Guest editorial: Multi-energy microgrid: Modelling, operation, planning, and energy trading. Energy Conversion and Economics, 2(3): 119–121, 2021.

[3] E. Guelpa, A. Bischi, V. Verda, M. Chertkov, and H. Lund.Towards future infrastructures for sustainable multi-energy systems: A review. Energy, 184: 2–21, 2019.

[4] F. Fang, Q.H. Wang, Y. Shi. A novel optimal operational strategy for the CCHP system based on two operating modes. IEEE Trans Power Syst, 27(2): 1032-1041, 2012.

[5] A.D. Smith, P.J. Mago Effects of load-following operational methods on combined heat and power system efficiency Appl Energy, 115: 337-351, 2014.

[6] N. Fumo, L.M. Chamra Analysis of combined cooling, heating, and power systems based on source primary energy consumption Appl Energy, 87(6): 2023-2030, 2010.

[7] N. Fumo, P.J. Mago, L.M. Chamra Emission operational strategy for combined cooling, heating, and power systems Appl Energy, 86(11): 2344-2350, 2009.

[8] Z. Li; Y. Xu; S. Fang; and X. Zheng. Robust Coordination of a Hybrid AC/DC Multi-Energy Ship Microgrid With Flexible Voyage and Thermal Loads. IEEE Transactions on Smart Grid, 11(4): 2782–2793, 2020.

[9] Z. Yang; J. Hu; and X. Ai, et al. Transactive Energy Supported Economic Operation for Multi-Energy Complementary Microgrids. IEEE Transactions on Smart Grid, 12(1): 4–17, 2020.

[10] C. A. Floudas and X. Lin. Mixed integer linear programming in process scheduling: Modeling algorithms and applications. Ann. Oper. Res., 139(1): 131-162, 2005.

[11] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers, 2010, doi: 10.1561/2200000016.

[12] H. Hua, J. Zhai, et al. Dynamic Partitioning and Energy Local Autonomy of Virtual Microgrid Groups Based on Enhanced Elite Preservation Genetic Algorithm. Proceedings of the CSEE: 1-15.

[13] H. Hua, J. Shi, X. Chen, et al. Carbon emission flow based energy routing strategy in energy Internet. IEEE Transactions on Industrial Informatics, 20(3): 3974-3985, 2024.

[14] H. Zhang, Q. Cao, H. Gao, P. Wang, W. Zhang and N. Yousefi. Optimum design of a multi-form energy hub by applying particle swarm optimization. J. Cleaner Prod., 260:121079, 2020.

[15] C. Timothée, A. T. D. Perera, J.-L. Scartezzini and D. Mauree. Optimum dispatch of a multi-storage and multi-energy hub with demand response and restricted grid interactions. Energy Proc., 142: 2864-2869, 2017.

[16] H. Hua, Y. Li, T. Wang, et al. Edge computing with artificial intelligence: A machine learning perspective. ACM Computing Surveys, 55(9): 1-35, 2023.

[17] H. Hua, Z. Qin, Y. Qin, et al. Data-driven dynamical control for bottom-up energy Internet system. IEEE Transactions on Sustainable Energy, 13(1): 315-327, 2022.

[18] P. Dai, W. Yu, G. Wen, et al. Distributed Reinforcement Learning Algorithm for Dynamic Economic Dispatch With Unknown Generation Cost Functions. IEEE Transactions on Industrial Informatics, 16(4): 2258-2267, 2020

[19] T. Yang, L. Zhao, W. Li, et al. Reinforcement learning in sustainable energy and electric systems: a survey," Annual Reviews in Control, 49(1): 145-163, 2020.

[20] T. Yang, L. Zhao, W. Li et al. Dynamic energy dispatch strategy for integrated energy system based on improved deep reinforcement learning. Energy. [Online]. pp. 1-15. Available: https://doi.org/10.1016/j.energy.2021.121377

[21] S. Fujimoto, H. V. Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. in Proc. 35th Int. Conf. Mach. Learn. (ICML), Stockholm, pp. 2587–2601, 2018.

[22] Z. Li and Y. Xu, Temporally-coordinated optimal operation of a multi-energy microgrid under diverse uncertainties. Appl. Energy, 240: 719–729, 2019.

[23] Ma. T, Wu. J, and L. Hao. Energy flow calculation and integrated simulation of micro-energy grid with combined cooling, heating and power[J]. Automat. Electric Power Syst., 40(23): 22–27, 2016.

[24] Z. Zhao, J. Guo, X. Luo, et al. Distributed robust model predictive control-based energy management strategy for islanded multi-microgrids considering uncertainty[J]. IEEE Transactions on Smart Grid, 13(03): 2107-2120, 2022.

[25] S. Wu, W. Hu, Z. Lu, Y. Gu, B. Tian and H. Li. Power system flow adjustment and sample generation based on deep reinforcement learning. J. Modern Power Syst. Clean Energy, 8(6): 1115-1127, 2020.

[26] Z. Li and Y. Xu. Dynamic dispatch of grid-connected multi-energy microgrids considering opportunity profit. 2017 IEEE Power & Energy Society General Meeting, Chicago, IL, USA, pp. 1-5, 2017, doi: 10.1109/PESGM.2017.8274205.