# Energy Optimization Management Scheme for Manufacturing Systems Based on BMAPPO: A Deep Reinforcement Learning Approach

Zhe Shao*
Woosong University, Endicott College, Korea, 34606

***Abstract*—To address the depletion of traditional energy sources and the increasingly severe environmental pollution, countries around the world have accelerated the deployment of renewable energy generation equipment. Energy optimization management for microgrids can address the randomness of factors such as renewable energy generation and load, ensuring the safe and stable operation of the system while achieving objectives such as cost minimization. Therefore, this paper conducts an in-depth study of energy optimization management schemes for microgrids and designs a multi-microgrid energy optimization management model and algorithm based on deep reinforcement learning. For the joint optimization problem among multiple microgrids with power flow between them, a two-layer energy optimization management scheme based on the multi-agent proximal policy optimization (PPO) algorithm and optimal power flow (BMAPPO) is proposed. This scheme is divided into two layers: first, the lower layer uses the multi-agent proximal policy optimization algorithm to determine the output of various controllable power devices in each microgrid; then, based on the lower layer's optimization results, the upper layer uses a second-order cone relaxation optimal power flow model to solve the optimal power flow between multiple microgrids, achieving power scheduling among them; finally, the total cost of the upper and lower layers is calculated to update the network parameters. Experimental results show that compared with other schemes, the proposed scheme achieves multi-microgrid energy optimization management at the lowest cost while ensuring online execution speed.**

***Keywords*—Microgrid; energy optimization management; deep reinforcement learning; multi-agent; Proximal Policy Optimization (PPO)***

## I. INTRODUCTION

Electric power is an indispensable driving force in modern society. In recent years, with the rapid development of technology and the continuous growth of the global population, the demand for electricity has been increasing year by year [1]. To meet this demand, the current smart grid is transitioning towards a more structured system based on microgrids. This transition, which optimizes energy storage systems through collaboration and self-organization, is key to driving the existing energy system towards being more intelligent, robust, and green.

Microgrid-based smart grids are not only better at integrating emerging distributed components but also position microgrids as an effective part of distribution and transmission system management through the evolving flexibility markets and new grid management concepts. The flexibility provided by microgrids, as well as their ability to operate in both grid-connected and island modes, are crucial solutions to the challenges faced by future transmission and distribution networks. The introduction of microgrids helps improve the reliability of power systems, reduce emissions, and expand the energy options for future power systems. Furthermore, the multi-microgrid structure formed by the interconnection of microgrids enhances the resilience, security, and intelligence of energy systems, supporting energy systems that incorporate large amounts of variable renewable energy.

Although the technologies related to components such as renewable energy generation and storage systems in microgrids have matured, joint optimization in microgrids remains challenging due to uncertainties in renewable energy generation, load, and energy prices. In addition to dealing with the fluctuations in renewable energy and load, storage devices must be optimized and controlled according to their operating costs or physical constraints. When multiple microgrids need to be optimized simultaneously, the complexity of the algorithms also increases. Given the technical and economic advantages of microgrids in future energy systems, ensuring the efficient and stable operation of microgrids has become a hot research topic.

A multi-microgrid refers to a system where multiple individual microgrids within a certain area are interconnected to achieve power mutual assistance. Compared to a single microgrid, a multi-microgrid has several advantages: First, a multi-microgrid can integrate large-scale renewable energy generation equipment, achieving a higher penetration rate of renewable energy through power flow between microgrids; second, it allows for the shared use of devices such as energy storage and generation, enabling a microgrid with large-capacity energy storage or high-power generation equipment to supply energy to other microgrids when the main grid's electricity price is high, further reducing costs; third, it enhances system robustness, allowing energy to be sourced from other microgrids if a microgrid's supply equipment fails or if the main grid experiences a power outage. Therefore, from the perspective of economic efficiency and future development trends, it is necessary to conduct research on energy optimization management for multi-microgrids.

Microgrid energy optimization management is an important research area in the power industry, aiming to achieve intelligent control and autonomous scheduling decisions for microgrids through optimization techniques. Under the premise

of ensuring the safe operation of equipment, the output of controllable power devices in microgrids is optimized to cope with fluctuations in renewable energy generation, load, and real-time electricity prices, meet load demand, avoid power waste, and minimize operational costs.

For the scenario of multiple microgrids, this paper proposes an energy optimization management scheme based on multi-agent proximal policy optimization. Since a single microgrid has limited capacity to deal with system uncertainties and often needs to trade with the main grid, joint optimization of multiple microgrids is expected to become a future development trend. To this end, this paper proposes a dual-layer structure for multi-microgrid energy optimization management. In the lower layer, the output of devices in each microgrid is decided based on a multi-agent proximal policy optimization algorithm, where centralized training ensures optimization effectiveness and decentralized decision-making protects user privacy. In the upper layer, the optimal power flow between microgrids is solved using a second-order cone relaxation optimal power flow model, ensuring power mutual assistance between microgrids.

The organization of this paper is as follows: Section II provides a detailed review of existing microgrid energy optimization management schemes; Section III proposes a dual-layer energy optimization management scheme based on multi-agent reinforcement learning to further improve the stability of microgrid operation and reduce costs; Section IV verifies the feasibility of the proposed method through case studies; Section V summarizes the contributions of the entire paper.

## II. Literature Review

This section will review existing work related to resource management to highlight the gaps in current research.

### A. Related Works

Energy optimization management in microgrids is essentially a constrained optimization problem with uncertain factors. Currently, the methods used in the field of microgrid energy optimization management mainly include metaheuristic algorithms (e.g., genetic algorithms), mathematical programming methods (e.g., mixed-integer linear programming), robust optimization, stochastic optimization, model predictive control, and deep reinforcement learning algorithms.

Metaheuristic algorithms have been widely applied in the field of microgrid energy optimization management. Among them, genetic algorithms and particle swarm optimization are the most commonly used, with similar techniques including ant colony optimization [2], crow search algorithm [3], and simulated annealing algorithm [4], among others. Torkan et al. [5] applied a multi-objective genetic algorithm to the optimization management of microgrids, considering the uncertainties brought by demand response (DR) programs, reactive power loads, and renewable energy. They optimized microgrid operations with energy cost and greenhouse gas emissions as objective functions, while this optimization objective function was constrained by a series of system constraints and was solved using a genetic algorithm.

In mathematical programming-based schemes, mixed-integer linear programming (MILP) can handle optimization problems where variables are continuous and discrete, making it highly suitable for application in microgrid energy optimization management. MILP can be used to establish mathematical models of microgrid components and optimize the cost function. Sigalo et al. [6] proposed an energy management scheme for grid-connected microgrids focused on battery storage systems, considering changes in grid electricity prices, renewable energy generation, and load demand, and determined the charging and discharging power of the battery to minimize the overall energy loss cost.

Stochastic and robust optimization-based microgrid energy optimization management schemes have been proposed to address the stochastic factors and prediction errors inherently present in microgrids. Chen et al. [7] proposed a new cumulative regret-based robust optimization method for the optimal management of grid-connected multi-energy microgrids considering uncertainty factors. Compared to traditional robust optimization methods, the proposed strategy ensures the robustness of microgrids and reduces the conservatism of microgrid operations. Additionally, by considering the demand response of thermal loads, the optimization model for microgrid energy management was improved. Abunima et al. [8] proposed a two-stage microgrid optimization scheduling method that coordinates microgrid assets under uncertainty, allowing microgrid operators to save operational costs without increasing investment costs, while meeting load demand. Nair et al. [9] considered an islanded microgrid composed of photovoltaic generation, supercapacitors, and regenerative fuel cells, utilizing a model predictive control algorithm. The goal was to enhance the utilization of renewable energy, improve microgrid operational efficiency, and reduce the degradation rate of the storage system.

A common feature of the above model-based methods is their reliance on precise predictions of uncertain factors in microgrids. Once prediction errors occur, the performance of these methods can be significantly impacted. Additionally, the computational cost of these methods is typically high, facing the issue of "curse of dimensionality"; as the complexity of the optimization problem increases, the computational cost multiplies, making it difficult to meet the real-time requirements of microgrid energy optimization management. To address these issues, some researchers have employed deep reinforcement learning (DRL) to solve the problem of microgrid energy optimization management. Deep reinforcement learning is a data-driven or model-free algorithm that does not rely on precise modeling of the microgrid environment. Thanks to the powerful perception capabilities of deep learning algorithms, DRL can effectively learn the microgrid environment model. Additionally, due to the strong decision-making ability of reinforcement learning, it can efficiently solve optimization problems. Alabdullah et al. [10] proposed a microgrid energy management solution based on the Deep Q-Network (DQN) algorithm, considering the stochastic behavior of various factors in the microgrid and modeling different grid components, while adhering to various power flow constraints in real-world environments.

### B. Research Gaps and Motivation

Based on the above literature review, the following conclusions can be drawn:

*1) Real-time performance and stability challenges:* Current microgrid energy optimization methods, including mathematical programming, stochastic optimization, robust optimization, and MPC, struggle with high computational complexity and real-time performance due to the handling of uncertainties. MPC, in particular, faces difficulties in ensuring stability and has not fully accounted for model uncertainties.

*2) Dependency on accurate predictions:* Model-based methods rely heavily on accurate predictions of uncertainties in microgrids. Prediction errors and high computational costs can lead to performance issues and the "curse of dimensionality," making real-time energy optimization challenging.

*3) Need for further research in deep reinforcement learning:* While deep reinforcement learning offers potential advantages such as real-time scheduling and a general framework, its applicability across different microgrid architectures and its effectiveness in online optimization require further research and validation.

Addressing these issues is critical for improving the effectiveness and applicability of microgrid energy management, which is the focus of this paper.

## III. Multi-Microgrid Dual-Layer Energy Optimization Management Model Based on Reinforcement Learning

### A. Multi-Microgrid Dual-Layer Energy Optimization Management Model

The structure of the multi-microgrid dual-layer energy optimization management model designed in this chapter is shown in Fig. 1. The lower layer consists of N microgrids, each containing photovoltaic generation equipment, wind power generation equipment, energy storage devices, micro gas turbines, loads, and a control center. The microgrids are interconnected through energy routers.

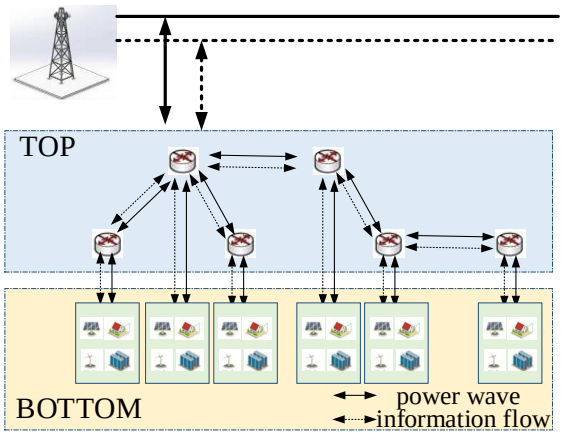


Fig. 1. Schematic diagram of the multi-microgrid dual-layer structure.

The upper layer is an abstracted topology based on the lower layer, which facilitates power flow analysis. In the process of one round of multi-microgrid energy optimization management, the lower layer first uses multi-agent deep reinforcement learning to decide the output of controllable power devices [11]. Then, the upper layer calculates the optimal power flow based on the regulated results of the lower layer. Finally, based on the results of the two-layer optimization, the reward value is calculated, and the neural network parameters are updated. Therefore, the following sections will first introduce the multi-agent deep reinforcement learning algorithm MAPPO in the lower layer, then explain the optimal power flow model for the multi-microgrid in the upper layer, and finally present the overall algorithm flow and experimental results.

### B. Lower Layer Multi-Agent Deep Reinforcement Learning Algorithm

As shown in Fig. 2, MAPPO is a multi-agent variant of PPO (Proximal Policy Optimization) and operates using a Centralized Training with Decentralized Execution (CTDE) approach. In centralized training, the Critic network of each agent can use global information during the offline training phase to achieve better convergence. In decentralized execution, each agent's Actor network can only observe its own state to make decisions during the online execution phase. Extensive experiments have shown that the clipped form of PPO consistently outperforms the penalized form, and hence, the clipped form is adopted in this work. The optimization objective can be written as follows:

$$\max \hat{\mathbb{E}}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_t\right)\right] \quad (1)$$

$$\text{clip}(r_t(\theta),1-\varepsilon,1+\varepsilon) = \begin{cases} 1-\varepsilon, & r_t(\theta) < 1-\varepsilon \\ 1+\varepsilon, & r_t(\theta) > 1+\varepsilon \\ r_t(\theta), & \text{other} \end{cases} \quad (2)$$

Here, clip(·) is the clipping function. When $\hat{A}_t > 0$, it indicates that the action a_t taken at this moment is better than the average, so maximizing Eq. (1) will increase $r_t(\theta)$, meaning the probability of action a_t in the new policy will increase. However, $r_t(\theta)$ will not increase beyond 1+ε. Conversely, when $\hat{A}_t < 0$, it indicates that the action $a_t$ taken at this moment is worse than the average, so maximizing Eq. (1) will decrease $r_t(\theta)$, meaning the probability of action a_t in the new policy will decrease. However, $r_t(\theta)$ will not decrease below 1-ε.

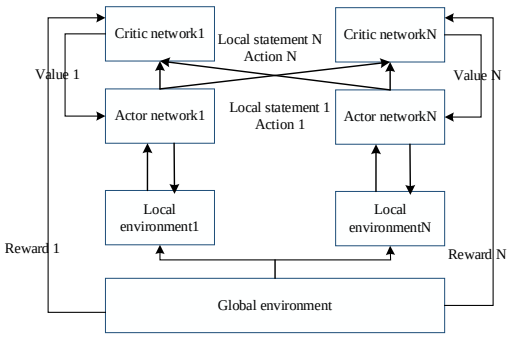


Fig. 2. MAPPO algorithm architecture.

### C. Upper-Layer Optimal Power Flow Model

Currently, distribution networks are primarily radial in structure. A radial structure with multiple microgrids can be modeled using the Branch Flow Model (BFM). Fig. 3 illustrates a schematic diagram of the Branch Flow Model. For node j:

- $V_j$ represents the voltage at the node;
- $s_j = p_j + iq_j$represents the power injection at the node.

For the branch from node i to node j (i→j):

- $I_{ij}$ represents the branch current;
- $S_{ij} = P_{ij} + iQ_{ij}$ represents the power at the sending end of the branch;
- $Z_{ij} = r_{ij} + ix_{ij}$ represents the branch impedance.

The topology of the upper-level multi-microgrid system is denoted as G(N,E)G(N,E). Finally, by applying angle relaxation and second-order cone relaxation, the optimal power flow problem is transformed into a convex optimization problem. Specifically:

$$l_{ij} \geq \frac{P_{ij}^2 + Q_{ij}^2}{v_i}, \forall (i,j) \in E \Leftrightarrow \left\| \begin{bmatrix} 2P_{ij} \\ 2Q_{ij} \\ l_{ij} - v_i \end{bmatrix} \right\|_2 \leq l_{ij} + v_i, \quad \forall (i,j) \in E \tag{3}$$

At this point, the optimization variables for the optimal power flow problem become $\{p_i, q_i, P_{ij}, Q_{ij}, l_{ij}, v_i\}$. For a radial network, the literature has proven that angle relaxation is tight; under the conditions that the objective function is strictly increasing and convex, the second-order cone relaxation is also tight. At this stage, the optimal power flow problem has been modeled as a convex optimization problem, which can be conveniently solved using commercial solvers like Gurobi.

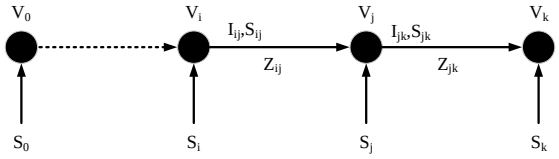


Fig. 3. Schematic diagram of branch flow structure.

### D. BMAPPO: Multi-Microgrid Dual-Layer Energy Optimization Management Algorithm

The constraints in a multi-microgrid system arise from three aspects: the electrical boundary constraints of various devices in the microgrid, the power balance constraints of each microgrid, and the power flow constraints between microgrids [12]. For convenience, unless otherwise stated, i represents the ith microgrid in the multi-microgrid system; t represents the tth round of optimization management; N represents the set of microgrid nodes; E represents the set of edges (power lines) between microgrids.

*1) Device boundary constraint*

*a) Micro gas turbine boundary constraints:*

$$P_{i,\min}^{MT} < P_{i,t}^{MT} < P_{i,\max}^{MT}, \forall i \in N \tag{4}$$

In the above formula, $P_{i,t}^{MT}$ is the output power of the micro gas turbine (kW); $P_{i,\min}^{MT}$ and $P_{i,\max}^{MT}$ are the lower and upper bounds of the micro gas turbine's output power (kW).

*b) Energy storage device boundary constraints:*

$$P_{i,\min}^{ESS} < P_{i,t}^{ESS} < P_{i,\max}^{ESS}, \forall i \in N \tag{5}$$

$$SOC_{i,\min} \leq SOC_{i,t} \leq SOC_{i,\max} \ \forall i \in N \tag{6}$$

In the above formulas, $P_{i,tS}^{ESS}$ is the charging/discharging power of the energy storage device (kW); $P_{i,\min}^{\mathrm{ESS}}$ and $P_{i,\max}^{ESS}$ are the lower and upper bounds of the charging/discharging power of the energy storage device (kW); $SOC_{i,t}$ is the state of charge of the energy storage device; $SOC_{i,\min}$ and $SOC_{i,\max}$ are the lower and upper bounds of the state of charge of the energy storage device.

*c) Main grid boundary constraints*

$$P_{i,\min}^{MG} < P_{i,t}^{MG} < P_{i,\max}^{MG}, \forall i \in N \tag{7}$$

$$Q_{i,\min}^{MG} < Q_{i,t}^{MG} < Q_{i,\max}^{MG}, \forall i \in N \tag{8}$$

In the above formulas, $P_{i,t}^{MG}$ is the active power traded between the microgrid and the main grid (kW); $P_{i,\min}^{MG}$ and $P_{i,\max}^{MG}$ are the lower and upper bounds of the active power traded between the microgrid and the main grid (kW); $Q_{i,t}^{MG}$ is the reactive power traded between the microgrid and the main grid (kVar); $Q_{i,\min}^{MG}$ and $Q_{i,\max}^{MG}$ are the lower and upper bounds of the reactive power traded between the microgrid and the main grid (kVar).

*2) Power balance constraints*

*a) Active power balance constraint*

$$p_{i,t} = P_{i,t}^{PV} + P_{i,t}^{WT} + P_{i,t}^{MT} + P_{i,t}^{ESS} + P_{i,t}^{MG} - P_{i,t}^{Load}, \forall i \in N \tag{9}$$

In the above formula, $p_{i,t}$ is the injected active power (kW); $P_{i,t}^{PV}$ and $P_{i,t}^{WT}$ are the active power outputs of photovoltaic and wind power generation, respectively (kW); $P_{i,t}^{MT}$ is the output power of the micro gas turbine (kW); $P_{i,t}^{ESS}$ is the charging/discharging power of the energy storage device (kW); $P_{i,t}^{MG}$ is the active power traded between the microgrid and the main grid (kW); $P_{i,t}^{\mathrm{Load}}$ is the active power of the load (kW).

*b) Reactive power balance constraint*

$$q_{i,t} = Q_{i,t}^{MG} - Q_{i,t}^{Lood}, \forall i \in N \tag{10}$$

In the above formula, $q_{i,t}$ is the injected reactive power (kVar); $Q_{i,t}^{\mathrm{MG}}$ is the reactive power from the main grid (kVar); $Q_{i,t}^{\mathrm{Load}}$ is the reactive power of the load (kVar).

*3) Power flow constraints*

$$v_j = v_i - 2\left(r_{ij}P_{ij} + x_{ij}Q_{ij}\right) + \left(r_{ij}^2 + x_{ij}^2\right)l_{ij}, \forall (i,j) \in E \tag{11}$$

$$p_j = \sum_{k:j\to k} P_{jk} - \sum_{i:i\to j} \left(P_{ij} - r_{ij}l_{ij}\right), \forall j \in N \tag{12}$$

$$q_j = \sum_{k:j\to k} Q_{jk} - \sum_{i:i\to j} \left(Q_{ij} - x_{ij}l_{ij}\right), \forall j \in N \tag{13}$$

$$l_{ij} \geq \frac{P_{ij}^2+Q_{ij}^2}{v_i},\ \forall(i,j) \in E \quad (14)$$

$$\left|I_{ij}\right| \leq \bar{I_{ij}},\ \forall(i,j) \in E \quad (15)$$

$$\underline{V_i} \leq |V_i| \leq \bar{V_i},\ \forall i \in N \quad (16)$$

$$\underline{s_i} \leq s_i \leq \bar{s_i},\ \forall i \in N \quad (17)$$

*E. Optimization Objective Function*

In this chapter, considering the uncertainties in renewable energy generation, load, and electricity prices, the objective is to minimize the cooperative operation cost of multiple microgrids. An energy optimization management model for multiple microgrids is constructed, and the objective function is as follows:

$$\min F_t = \min \sum_{t=1}^{T} \left(\sum_{i=1}^{N} \left(F_{i,t}^{MG} + F_{i,t}^{MT} + F_{i,t}^{ESS}\right) + F_t^{\text{Loss}}\right) \quad (18)$$

Where:

- $F_t$ is the total operating cost of the multi-microgrid system.
- t represents the t-th round of optimization management.
- T is the total number of rounds within an energy optimization management period.
- $F_{i,t}^{MG}$ represents the cost of trading with the main grid.
- $F_{i,t}^{MT}$ represents the generation cost of the micro gas turbine.
- $F_{i,t}^{ESS}$ represents the loss cost of the energy storage device.
- $F_t^{\text{Loss}}$ represents the power transmission loss between microgrids.

For simplicity, in this chapter, ΔtΔt represents the time length of one round of optimization management (hours).

*a) Cost of trading between microgrid and main grid*

$$F_{i,t}^{MG} = c_t^{MG} P_{i,t}^{MG} \cdot \Delta t \quad (19)$$

In the above formula, ctMGctMG is the electricity price of the main grid during the tt-th round of optimization management ($/kWh), and P(i,t)MGP(i,t)MG is the active power purchased by the microgrid from the main grid (kW).

*b) Micro gas turbine generation cost*

$$F_{i,t}^{MT} = \left(a \cdot \left(P_{i,t}^{MT}\right)^2 + b \cdot P_{i,t}^{MT} + c\right) \cdot \Delta t \quad (20)$$

In the above formula, aa, bb, and cc are cost coefficients.

*c) Energy storage device loss cost*

$$F_{i,t}^{ESS} = \left(c^{\text{ESS}} \cdot \left(P_{i,t}^{cha} \cdot \eta_{\text{cha}} + P_{i,t}^{dis} / \eta_{dis}\right)\right) \cdot \Delta t \quad (21)$$

In the above formula, cESScESS is the loss coefficient; P(i,t)chaP(i,t)cha and P(i,t)disP(i,t)dis are the charging and discharging powers of the energy storage device, respectively (kW); ηchaηcha and ηdisηdis are the charging and discharging efficiencies of the energy storage device, respectively.

*d) Transmission loss between microgrids*

$$F_t^{\text{Loss}} = \sum_{ij \in E} c_t^{MG} \left|I_{ij,t}\right|^2 r_{ij} \cdot \Delta t \quad (22)$$

In the above formula, $I_{ij_{,,}}$ and $r_{ij}$ are the current (A) and impedance (Ω) between microgrid i and microgrid j, respectively.

Based on the above description, solving the optimization problem in Eq. (19) requires handling optimization variables that can be divided into lower-layer and upper-layer optimization variables. The lower-layer optimization variables include:

- The generation power of micro gas turbines in microgrids $\{P_{i,t}^{MT}\}$.
- The charging and discharging power of energy storage devices in microgrids $\{P_{i,t}^{ESS}\}$.
- The active power traded between the microgrid and the main grid $\{P_{i,t}^{\text{MG}}\}$.
- The reactive power traded between the microgrid and the main grid $\{Q_{i,t}^{\text{MG}}\}$.

The upper-layer optimization variables include:

- Node voltage $\{v_{j,t}\}$.
- Node injected power $\{p_{j,t}, q_{j,t}\}$.
- Branch current $\{l_{ij,t}\}$ and branch power $\{P_{ij,t}, Q_{ij,t}\}$.

*F. Construction of a Partially Observable Markov Decision Process*

The multi-agent reinforcement learning algorithm MAPPO (Multi-Agent Proximal Policy Optimization) is used to solve the multi-microgrid energy optimization management problem, which can be modeled as a Partially Observable Markov Decision Process (POMDP). POMDP can be defined as a five-tuple $(\mathcal{S}, \{\mathcal{O}_i\}_{i=1}^n, \{\mathcal{A}_i\}_{i=1}^n, \mathcal{P}, \{r_i\}_{i=1}^n)$, where:

- $\mathcal{S}$ is the global state space.
- $\mathcal{P}$ is the state transition function.
- For agent $i$, the observation space is $\mathcal{O}_i$, the action space is $\mathcal{A}_i$, and the reward function is $r_i$.

Further, in multi-agent reinforcement learning, the interaction process between agents and the environment is as follows:

- At time step t, each agent i obtains an observation state $o_{i,t} \in \mathcal{O}_i$ and selects an action $a_{i,t} \in \mathcal{A}_i$ according to the policy $\pi_i: \mathcal{O}_i \times \mathcal{A}_i \rightarrow [0,1]$.
- Then, the system transitions to the next state $o_{i,t+1}$ according to the state transition probability $\mathcal{P}$ and receives a reward $(s_t, a_{i,t})$.

The objective of multi-agent reinforcement learning is to maximize the cumulative return: $J(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \frac{1}{n} \gamma^t \sum_{i=1}^{n} r_{i,t}\right]$.

Therefore, the multi-microgrid energy optimization management problem is modeled as a POMDP below.

*a) Observation space definition*

$$o_{i,t} = \left[P_{i,t}^{PV}, P_{i,t}^{WT}, P_{i,t}^{Load}, c_t^{MG}, SOC_{i,t}\right] \quad (23)$$

In the above formula:

- $P_{i,t}^{PV}$ and $P_{i,t}^{WT}$ are the photovoltaic and wind power generation outputs, respectively.
- $P_{i,t}^{Load}$ is the load power.
- $c_t^{MG}$ is the real-time electricity price of the main grid.
- $SOC_{i,t}$ is the state of charge of the energy storage device.

*b)* b) State Space Definition

$$s_t = \left[o_{1,t}, o_{2,t}\ \mathrm{K}, o_{n,t}\right] \quad (24)$$

The state space contains the observation space of all agents and represents the global information of the multi-agent environment.

*c) Action space definition*

$$a_{i,t} = \left[P_{i,t}^{MT}, P_{i,t}^{ESS}, P_{i,t}^{\mathrm{MG}}, Q_{i,t}^{\mathrm{MG}}\right] \quad (25)$$

In the above formula:

- $P_{i,t}^{MT}$ is the power output of the micro gas turbine.
- $P_{i,t}^{ESS}$ is the charging/discharging power of the energy storage device.
- $P_{i,t}^{\mathrm{MG}}$ is the active power traded between the microgrid and the main grid.
- $Q_{i,t}^{\mathrm{MG}}$ is the reactive power traded between the microgrid and the main grid.

*d) Reward function definition*

$$r_{i,t} = -\left(\alpha \cdot \left(F_{i,t}^{MG} + F_{i,t}^{MT} + F_{i,t}^{ESS}\right) + \beta \cdot \sum_{i=1}^{N} \mathrm{cons}\left(SOC_{i,t}\right)\right) \quad (26)$$

$$r_t = -F_t^{\mathrm{Loss}} + \sum_{i=1}^{N} r_{i,t} \quad (27)$$

Eq. (26) is the reward function for agent ii, where α and β are coefficients, and cons (·) is a penalty function introduced in the previous chapter. Eq. (27) is the total reward for the multi-microgrid system. Since the algorithm designed in this chapter is a dual-layer optimization, the total reward is calculated after the upper-layer optimization is completed. It includes the sum of all lower-layer agent rewards and the upper-layer optimization objective $F_t^{\mathrm{Loss}}$. It is worth mentioning that Eq. (23) can also be decomposed using the vectorization approach proposed in the previous chapter, but for simplicity, the traditional reward function is used here. Online Decision-Making Process of the MAPPO-Based Multi-Microgrid Dual-Layer Energy Optimization Management Scheme. The online decision-making process is shown in Algorithm 1.

**Algorithm 1:** Multi-Microgrid Energy Optimization Management Algorithm Online Decision-Making Process

**Input**: Actor network weights θπθπ

**Output**: Values of all optimization variables in the tt-th round of multi-microgrid energy optimization

Step 1: Obtain the initial state of all agents.

For t=1,2,…,T do:

Step 1: Each agent i∈N in the lower layer observes state $o_t^i$ and selects an action $a_{i,t}$ based on the policy, obtaining the optimization variables $\left\{P_{i,t}^{MT}, P_{i,t}^{ESS}, P_{i,t}^{\mathrm{MG}}, Q_{i,t}^{\mathrm{MG}}\right\}, i \in N$.

Step 2: The upper layer solves the optimal power flow problem using a commercial optimizer, obtaining the optimization variables $\left\{v_{j,t}\right\} \setminus \left\{p_{j,t}, q_{j,t}\right\} \setminus \left\{\Delta P_{i,t}^{MG}, \Delta Q_{i,t}^{MG}\right\}, i \in N$ and $\left\{l_{ij,t}\right\} \setminus \left\{P_{ij,t}, Q_{ij,t}\right\}, ij \in E$.

Step 3: Apply all optimization variables to the multi-microgrid system and obtain the next state $s_{t+1}$.

End for

## IV. Experimental Results Analysis

### A. Simulation Environment and Parameter Settings

The IEEE 33-bus system structure, as shown in Fig. 4, is used as the upper-level topology for the dual-layer energy optimization management of multiple microgrids. Microgrids can be placed at any node, and the impedance value between two adjacent nodes is the average of the branch impedances between the nodes. For example, if microgrids are connected at nodes 1 and 21 to form a system with two microgrids, the impedance between nodes 1 and 21 would be the average impedance of branches 1-18, 18-19, 19-20, and 20-21. This setup allows the generation of various simulation environments on the IEEE 33-bus system. In this chapter, nodes 1, 2, 5, and 24 are selected to connect microgrids 1 to 4, forming a multi-microgrid structure with four microgrids. Energy optimization management is performed every hour.

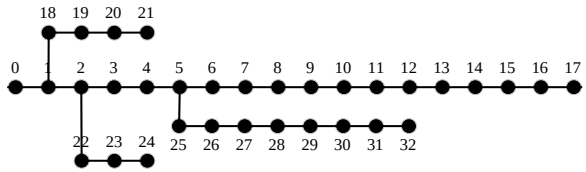


Fig. 4. Schematic diagram of the IEEE 33-bus system structure.

The experimental equipment used in this chapter includes an Intel(R) Core(TM) i5-10210U CPU @1.60GHz 2.10 GHz and an NVIDIA GeForce RTX2060. The compiler used is Pycharm 2022.3, and the programming language is Python 3.8. The commercial optimizer used for solving the upper-level optimal power flow is the Python version of Gurobi 10.0.1. The implementation framework for the lower-level multi-agent deep reinforcement learning algorithm is the mainstream neural network development framework Pytorch. The MAPPO parameters are set as shown in Table I:

TABLE I. MAPPO PARAMETER SETTINGS

| Parameter | Value |
|---|---|
| Discount Factor γ | 0.95 |
| Number of Neurons in Hidden Layer | 128 |
| Actor Network Learning Rate $l_a$ | 0.0003 |
| Critic Network Learning Rate $l_c$ | 0.0001 |
| Clipping Function Hyperparameter $\varepsilon$ | 0.2 |
| Size of Experience Replay Pool D | 10000 |
| Mini-batch Size B | 96 |
| Maximum Number of Training Epochs | 20000 |

### B. Optimization Management Results Analysis

*1) Convergence analysis:* As shown in Fig. 5, during the offline training process, the BMAPPO algorithm proposed in this chapter converges at around 9000 iterations. The smoothed cumulative reward oscillates around -2250. Since the reward function is a negative function of the optimization management cost, the convergence curve shows an upward trend, indicating that during the iteration process, the agent learns effective strategies to reduce the cost of multi-microgrid optimization management. During offline training, 80% of the dataset was used. Next, the network parameters after training are saved, and the remaining 20% of the dataset is used to simulate the online optimization management process to test the effectiveness of the algorithm.

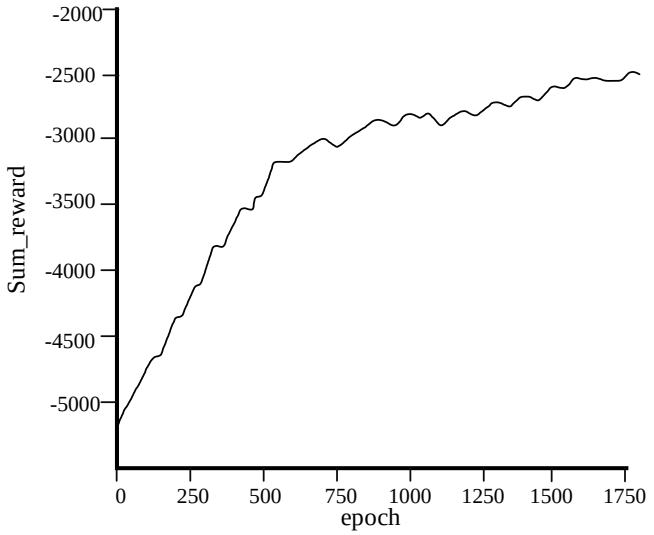


Fig. 5. Convergence curve of the BMAPPO algorithm during offline training.

*2) Effectiveness analysis:* Fig. 6 shows the lower-level optimization results for Microgrid 1 and Microgrid 2.

Fig. 6(a1) and 6(a2) represent the output curves of controllable power devices in Microgrid 1 and Microgrid 2, respectively, including the power generation of micro gas turbines, the charging and discharging power of energy storage devices, and the power regulation results of transactions with the main grid. Combined with the real-time electricity price fluctuations of the main grid shown in Fig. 5, it can be seen that when the main grid electricity price is low from 1 to 5 hours, the power of the energy storage devices in both microgrids is negative (except for a discharge of about 40 kW in Microgrid 1 at 5 hours), representing charging of the energy storage devices. During this period, a large amount of electricity is purchased from the main grid, while the micro gas turbines almost do not output power. When the main grid price is moderately high from 9 to 17 hours, the trend of output from controllable power devices is similar to that from 1 to 5 hours, but the load power gap in this period is also relatively low, so the power purchased from the main grid is also relatively less. When the electricity price is high from 6 to 8 hours and 18 to 22 hours, the output of the energy storage devices is generally positive, representing discharging of the energy storage devices. During these periods, the power purchased from the main grid is at a low point, and the micro gas turbines output a large amount of power to fill the load power gap. Therefore, the BMAPPO proposed in this chapter is effective in cost savings.

Fig. 6(b1) and 6(b2) show the state of charge (SoC) curves of the energy storage devices in Microgrid 1 and Microgrid 2, respectively, with the upper and lower bounds of the SoC indicated, set at 0.9 and 0.1. An SoC less than 0.1 indicates an over-discharging condition, while an SoC greater than 0.9 indicates an over-charging condition. It can be seen that the energy storage device in Microgrid 2 is in a safe state during the typical day, with no over-charging or over-discharging occurring. However, Microgrid 1 shows a slight over-charging condition from 16 to 19 hours.

Fig. 7 illustrates the upper-level optimization results for Microgrid 1 and Microgrid 2.

Fig. 7(a1) and 7(a2) display the power injection curves for Microgrid 1 and Microgrid 2, respectively. Positive power injection indicates that there is excess power within the microgrid that is not being consumed, while negative power injection indicates that there is an unsatisfied power deficit within the microgrid. The upper-level power flow optimization will balance the power injection between microgrids, ensuring power balance in each microgrid. From Fig. 7(a1), it can be seen that Microgrid 1 has positive power injection from 17 to 21 hours, which can be transmitted to other microgrids. From Fig. 7(a2), it can be seen that Microgrid 2 has negative power injection from 18 to 22 hours, which can be obtained from other microgrids. This shows that the upper-level microgrids can achieve mutual support.

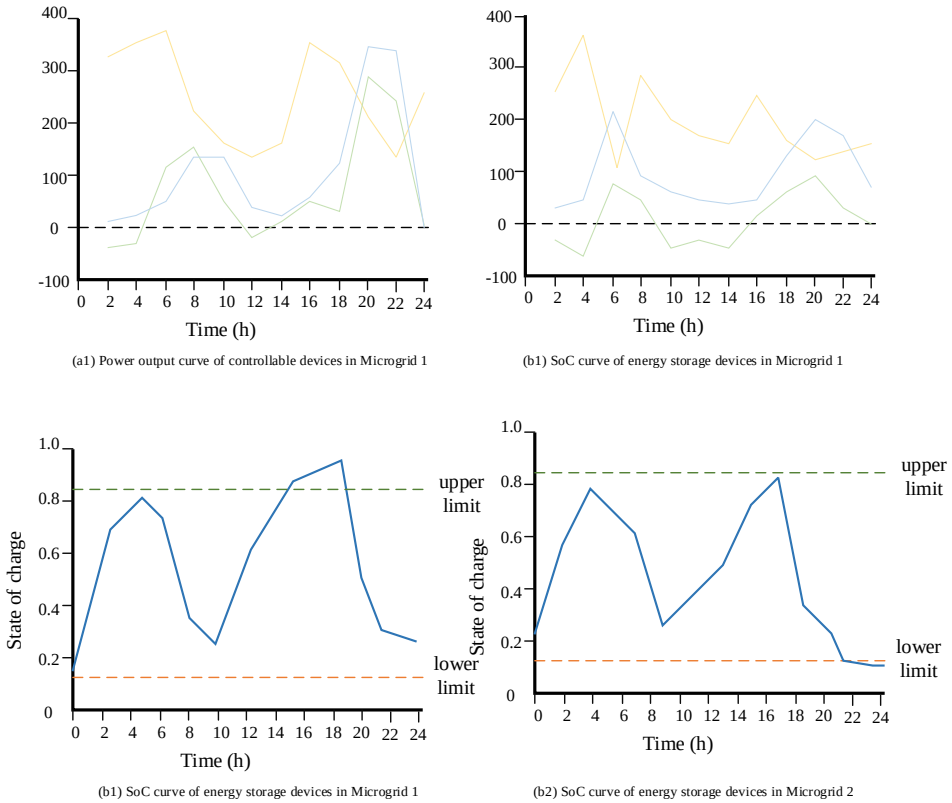


Fig. 6. Lower-level optimization results of Microgrid 1 and Microgrid 2.


Fig. 7. Upper-level optimization results of Microgrid 1 and Microgrid 2.

## V. CONCLUSION

This paper proposes a dual-layer optimization management scheme based on the multi-agent reinforcement learning algorithm MAPPO for the energy optimization management problem of multiple microgrids. The lower layer uses MAPPO to make decisions on the power output of each microgrid device, handling power imbalances, while the upper layer achieves overall power balance of multiple microgrids through a second-order cone relaxation optimal power flow model. The experimental results show that the designed BMAPPO algorithm effectively achieves mutual support between microgrids and significantly reduces the energy optimization management costs of multiple microgrids.

Although the deep reinforcement learning-based energy optimization management scheme proposed in this paper shows significant advantages in cost savings, there is still room for improvement:

The deep reinforcement learning method relies on a large amount of historical data for training, which may be difficult to obtain in practical applications. Therefore, future research should focus on reducing dependence on historical data or improving data utilization efficiency.

This paper only studies the situation where microgrids are connected to the main grid, while islanded microgrids have widespread applications in remote areas, where their optimization management cost control is of significant importance. Therefore, how to reduce the management costs of islanded microgrids while ensuring safety is an important direction for future research.

## ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Avoid the stilted expression, “One of us (R. B. G.) thanks . . .” Instead, try “R. B. G. thanks.”

## REFERENCES

[1] Fu T, Liu S, Li P. Intelligent smelting process, management system: Efficient and intelligent management strategy by incorporating large language model[J]. Frontiers of Engineering Management, 2024: 1-17.

[2] Liu S, Zheng P, Shang S. A novel bionic decision-making mechanism for digital twin-based manufacturing system[J]. Manufacturing Letters, 2023, 35: 127-131.

[3] Liu S, Zheng P. A Novel Bionic Digital Twin-Based Manufacturing System Toward the Mass Customization Paradigm[C]//2023 IEEE 19th International Conference on Automation Science and Engineering (CASE). IEEE, 2023: 1-6.

[4] ANGELIM J, AFFONSO C. Energy management on university campus with photovoltaic generation and BESS using simulated annealing[C]. proceedings of the 2018 IEEE Texas Power and Energy Conference (TPEC). IEEE, 2018: 1-6.

[5] TORKAN R, ILINCA A, GHORBANZADEH M. A genetic algorithm optimization approach for smart energy management of microgrids [J]. Renewable Energy, 2022, 197: 852-63.

[6] SIGALO M B, PILLAI A C, DAS S, et al. An energy management system for the control of battery storage in a grid-connected microgrid using mixed integer linear programming [J]. Energies, 2021, 14(19): 6212.

[7] CHEN T, CAO Y, QING X, et al. Multi-energy microgrid robust energy management with a novel decision-making strategy [J]. Energy, 2022, 239: 121840.

[8] ABUNIMA H, PARK W-H, GLICK M B, et al. Two-Stage stochastic optimization for operating a Renewable-Based Microgrid [J]. Applied Energy, 2022, 325: 119848.

[9] NAIR U R, COSTA-CASTELLó R. A model predictive control-based energy management scheme for hybrid storage system in islanded microgrids [J]. IEEE access, 2020, 8: 97809-22.

[10] ALABDULLAH M H, ABIDO M A. Microgrid energy management using deep Q-network reinforcement learning [J]. Alexandria Engineering Journal, 2022, 61(11): 9069-78.

[11] Fu T, Li P, Liu S. An imbalanced small sample slab defect recognition method based on image generation[J]. Journal of Manufacturing Processes, 2024, 118: 376-388.

[12] Fu T, Liu S, Li P. Digital twin-driven smelting process management method for converter steelmaking[J]. Journal of Intelligent Manufacturing, 2024: 1-17.