参赛队员姓名： 杨静姝

中学： 上海市浦东新区民办万科学校

省份： 上海

国家/地区： 中国，上海

指导教师姓名： 闻海洋

指导教师单位： 上海市浦东新区民办万科学校

论文题目： Research on Deep Reinforcement Learning Optimization for Collaborative Scheduling and Trading Mechanisms in Multi-Microgrid Systems under the Dual-Carbon Goals

2025 S.T.Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

Research on Deep Reinforcement Learning Optimization for Collaborative Scheduling and Trading Mechanisms in Multi-Microgrid Systems under the Dual-Carbon Goals

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

Abstract

Under the dual-carbon goals, the integration of high-penetration renewable energy into multi-microgrid systems introduces significant challenges, including increased uncertainty in supply and demand, inefficient coordination, and the lack of market-oriented trading mechanisms. To address these issues and achieve a balance between economic efficiency and operational stability, this study proposes a collaborative scheduling and energy trading framework that integrates deep reinforcement learning with Stackelberg game theory. A hierarchical optimization model is developed, where the distribution system operator (DSO) acts as the leader setting energy prices, and individual microgrids serve as followers optimizing their local operations in response to price signals. The microgrid decision-making process is formulated as a Markov decision process, incorporating renewable generation, load demand, energy storage, and demand response, while ensuring physical and operational constraints are satisfied. To enhance learning stability and accuracy in continuous action spaces, a PER-Dueling-DDQN architecture is employed for value-based learning, and the DDPG algorithm is adopted to solve the bilevel game, leveraging actor-critic networks and experience replay for efficient training. The distributed design ensures that sensitive operational data remain localized, preserving privacy while enabling coordinated control. Simulation results on an IEEE 33-node system demonstrate that the proposed approach achieves superior economic performance and robustness under uncertain conditions, effectively supporting market-driven, privacy-preserving coordination in multi-microgrid systems with high renewable penetration.

Keywords: Multi-Microgrid System; Dual-Carbon Goals; Deep Reinforcement Learning; Stackelberg Game; Collaborative Scheduling; Energy Trading Mechanism; PER-Dueling-DDQN; DDPG Algorithm

Contents

1 Introduction **4**
- 1.1 Background . . . 4
- 1.2 Literature Review . . . 4

2 List of Symbols **5**

3 Optimization Model for Multi-Microgrid Energy Management **6**
- 3.1 Objective Function . . . 6
- 3.2 Component Cost Models . . . 6
- 3.3 Constraints . . . 7

4 Markov Decision Process Framework for Microgrid Optimization **9**
- 4.1 MDP Formulation . . . 9
- 4.2 Objectives . . . 11
  - 4.2.1 Cumulative Reward . . . 11
  - 4.2.2 Value Functions . . . 11

5 Solution Methodology **11**
- 5.1 Neural Network Architecture . . . 12
- 5.2 Double Q-Network Design . . . 12
- 5.3 Loss Function . . . 13
- 5.4 Analysis of Overestimation in Deep Q-Networks . . . 13
- 5.5 Proposed Framework: PER-Dueling-DDQN . . . 13
  - 5.5.1 Dueling Network Architecture . . . 14
  - 5.5.2 Prioritized Experience Replay (PER) . . . 15
- 5.6 Training Process . . . 16

6 Coordinated Optimal Scheduling Based on Stackelberg Game and DDPG Algorithm **17**
- 6.1 System Architecture . . . 17
- 6.2 Upper-Level Optimization Model . . . 19
- 6.3 Lower-Level Optimization Model . . . 20
- 6.4 Network Constraints . . . 22
- 6.5 Stackelberg Game Framework . . . 22
- 6.6 Solution via Policy-Based Deep Reinforcement Learning . . . 25

7 Simulation Verification **30**
- 7.1 Simulation Parameter Settings . . . 30
- 7.2 Training Process Analysis . . . 30
- 7.3 Online Operation Performance . . . 31

8 Online Operation Analysis **35**

9 Discussion **36**

10 Conclusion **37**

11 Acknowledgements **39**

2025 S.T. Yau High School Science Award 仅用于2025丘成桐中学科学奖论文公示

1 Introduction

1.1 Background

Energy is a crucial foundation for economic and social development, with fossil fuels serving as the primary conventional source for power generation [1]. However, the large-scale extraction and consumption of fossil fuels have led to increasingly prominent issues of energy scarcity. Meanwhile, the combustion of fossil fuels causes environmental pollution and the greenhouse effect, severely impacting human health and socioeconomic development. Therefore, countries around the world are actively adjusting their energy structures, promoting the development of renewable energy to facilitate sustainable economic and social progress. An individual microgrid has limited capacity and limited self-regulation capability. If there are sudden fluctuations in renewable generation or load power that exceed its regulation range, the voltage and frequency stability of the microgrid may be compromised [2]. To address this issue, several geographically adjacent microgrids can be integrated into a multi-microgrid (MMG) system, enabling energy sharing and mutual support among them. This enhances the utilization efficiency of renewable energy and improves the overall system's operational reliability and stability. Energy exchange among multiple microgrids is often accompanied by market transactions. Under appropriate market mechanisms, the trading benefits within a multi-microgrid system are generally higher than those achieved by individual microgrids operating independently.

1.2 Literature Review

Recent studies have explored diverse approaches to optimize microgrid operation and energy management under increasing penetration of renewable sources and distributed energy systems. Parisio et al. [3] formulated the microgrid optimization problem as a mixed-integer linear program (MILP), incorporating both discrete variables—such as generator on/off states—and continuous variables like power flows, while enforcing system constraints. By leveraging commercial solvers such as CPLEX or Gurobi, their model predictive control strategy achieves near-optimal performance with sub-second computation times, enabling real-time implementation. In a related effort, Shayan et al. [4] applied a dynamic decision algorithm combined with a Markov-based prediction chain to optimize multi-microgrid systems under boost voltage control, demonstrating through 13 case studies that higher renewable penetration reduces fossil fuel consumption but may compromise economic viability unless supported by favorable feed-in tariffs—specifically above 0.06 USD/kWh for full renewable feasibility. To address the complexity of real-time scheduling in microgrids with multiple battery energy storage systems (BESSs), Shuai et al. [5] proposed an online optimization framework using a Branching Dueling Q-Network (BDQ), a deep reinforcement learning (DRL) method that effectively handles high-dimensional action spaces and scales well with system size, thus mitigating the curse of dimensionality. Complementing this, Li et al. [6] developed a federated dueling deep Q-network (DDQN) within an edge-cloud computing architecture to preserve data privacy and reduce communication overhead in distributed microgrid energy management, introducing a novel action exploration mechanism to enhance economic performance without sacrificing scalability. Game-theoretic approaches have also gained traction: Dong et al. [7] designed a hierarchical Stackelberg game model integrated with a multi-agent system for microgrid clusters, where the cluster operator sets incentive prices, individual microgrids optimize energy transactions, and end users adjust consumption accordingly, resulting in improved coordination and economic benefits across the network. Similarly, Liu et al. [8] employed a Stackelberg game framework for microgrids with photovoltaic (PV) prosumers, in which the microgrid operator acts as the leader by setting optimal prices, while prosumers respond as followers to maximize their utility, with a

2025 S.T. Yau High School Science Award
仅用于2025年成府中学科学论文公示

dedicated billing mechanism addressing uncertainties in PV generation and load demand; real-world data validated the model's ability to improve operator profits, prosumer satisfaction, and overall energy balance. Extending this concept, Li et al. [9] introduced a real-time pricing model based on an improved Stackelberg game for microgrids equipped with wind, solar, and energy storage systems, where joint optimization of storage scheduling and dynamic pricing led to a 31.89% increase in daily profits compared to unoptimized scenarios and a 5.4% gain over conventional methods. Finally, Hu et al. [10] integrated reinforcement learning with myopic optimization in a soft actor-critic DRL framework for multi-timescale microgrid coordination, where an actor-critic agent determines storage actions and a myopic model refines power flow decisions using real-time measurements processed through deep neural networks, achieving a 90.98% improvement in online energy management efficiency over the myopic approach alone.

2 List of Symbols

Table 1: Symbol Definitions

Symbol	Description	Unit
Δt	Time step index	h
$C_{cost}(t)$	Total operational cost of the microgrid system over the scheduling horizon at time t	¥
$C_{buy}(t)$	Cost of purchasing electricity from the main grid at time t	¥
$C_{sell}(t)$	Revenue from selling electricity to the main grid at time t	¥
$C_{MT}(t)$	Total operating costs of microturbines at time t	¥
$C_{DR}(t)$	Compensation cost for demand response interruptions at time t	¥
$C_{ESS}(t)$	Operating costs of energy storage systems at time t	¥
$C_{EX}(t)$	Cost/revenue associated with electricity exchanged between the microgrid and external grid (positive for purchase, negative for sale) at time t	¥
C_{ESS}	Unit charge/discharge cost (including battery maintenance and lifetime degradation)	¥
C_{DR}	Unit load curtailment compensation cost	\$ / kW
$P_{buy}(t)$	Power purchased from main grid at time t	kW
$P_{sell}(t)$	Power sold to main grid at time t	kW
$P_{load}(t)$	Total active power demand of residential appliances at time t	kW
$P_{WT}(t)$	Active power output of a wind turbine at time t	kW
$P_{PV}(t)$	Active power output of the photovoltaic system at time t	kW
$P_{DR}(t)$	Amount of load voluntarily or contractually reduced by consumers during DR events	kW
η_{cha}	Charging efficiency	/
η_{dis}	Discharging efficiency	/
a	Time-invariant costs of microturbine operation, including scheduled maintenance fees, start-up/shutdown costs...	/
b	The variable operating cost coefficient	/
SOC(t)	State of charge of the battery at time t	%

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

3 Optimization Model for Multi-Microgrid Energy Management

To enable cost-efficient and reliable operation of multi-microgrid systems under high renewable penetration, an optimization framework is formulated to minimize total operational costs over a defined scheduling horizon. The objective function integrates key components including grid power exchange, distributed generation, energy storage operations, and demand-side flexibility. This section presents the mathematical formulation of the cost minimization problem, along with auxiliary equations that define each cost component in detail.

3.1 Objective Function

The primary goal is to minimize the total operational cost across all time intervals within the scheduling period. This is expressed as:

$$\min \mathcal{C}_{\text{cost}}(t) = \mathcal{C}_{EX}(t) + \mathcal{C}_{MT}(t) + \mathcal{C}_{ESS}(t) + \mathcal{C}_{DR}(t), \tag{1}$$

where $\mathcal{C}_{\text{cost}}(t)$ denotes the total system cost at time t, composed of four major terms: $\mathcal{C}_{EX}(t)$ represents the net cost of electricity exchanged with the main grid; $\mathcal{C}_{MT}(t)$ accounts for the operating cost of microturbines; $\mathcal{C}_{ESS}(t)$ captures the operational expenses of battery energy storage systems; and $\mathcal{C}_{DR}(t)$ quantifies compensation paid to users participating in demand response programs. Each term is modeled to reflect real-world economic and physical constraints, ensuring practical applicability of the solution.

3.2 Component Cost Models

3.2.1 Grid Power Exchange Cost

Electricity trading between the microgrid cluster and the upstream grid involves both purchasing and selling actions. The associated cost or revenue is calculated as:

$$\mathcal{C}_{EX}(t) = (C_{buy}(t) \cdot P_{buy}(t) - C_{sell}(t) \cdot P_{sell}(t)) \cdot \Delta t, \tag{2}$$

where $C_{buy}(t)$ and $C_{sell}(t)$ are the time-varying electricity prices for buying and selling, respectively; $P_{buy}(t)$ and $P_{sell}(t)$ denote the imported and exported active power at time t; and Δt is the duration of the time step in hours. Note that power export is limited by contractual agreements and local regulatory policies, which are enforced through additional constraints in the model.

3.2.2 Microturbine Operating Cost

The operating cost of microturbines includes fixed start-up and maintenance charges, as well as variable fuel-related expenses dependent on output power. It is modeled as a linear function of generated power:

$$\mathcal{C}_{MT}(t) = (a + b \cdot P_{MT}(t)) \cdot \Delta t, \tag{3}$$

where a represents the fixed cost per hour regardless of generation level, b is the fuel cost coefficient in ¥/kW, and $P_{MT}(t)$ is the real power output of the microturbine at time t. This simplified quadratic-to-linear approximation balances accuracy and computational tractability, especially when embedded in mixed-integer programming frameworks.

2025 S. T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

3.2.3 Energy Storage System Cost

Battery degradation and maintenance contribute significantly to long-term operational expenses. The cost associated with charging and discharging cycles is given by:

$$\mathcal{C}_{ESS}(t) = C_{ESS} \cdot (P_{cha}(t) \cdot \eta_{cha} + P_{dis}(t) \cdot \eta_{dis}) \cdot \Delta t, \tag{4}$$

where C_{ESS} is the unit cost per effective power throughput (¥/kWh), $P_{cha}(t)$ and $P_{dis}(t)$ are the charging and discharging power levels, and η_{cha}, η_{dis} represent the corresponding charge and discharge efficiencies. This expression approximates aging effects based on round-trip energy flow, commonly used in short-term dispatch models.

3.2.4 Demand Response Compensation

Incentive-based demand response programs require financial compensation for load curtailment. The resulting cost is:

$$\mathcal{C}_{DR}(t) = C_{DR} \cdot P_{DR}(t) \cdot \Delta t, \tag{5}$$

where C_{DR} is the agreed-upon compensation rate (¥/kW), and $P_{DR}(t)$ is the amount of load reduced during event periods. This term encourages consumer participation while maintaining budgetary control within the overall optimization framework.

3.3 Constraints

This section presents the constraints imposed on the system to ensure its safe, stable, and economical operation. The constraints are designed to maintain electrical safety by preventing overloads and faults, guarantee system stability by satisfying operational limits, and promote economic efficiency by adhering to demand response and operational cost requirements. Together, these constraints define the feasible operating region of the system under various conditions.

3.3.1 Power Balance Constraint

At every time step, total generation and imports must match total demand and exports to ensure system stability. This can be expressed as:

$$P_{MT}(t) + P_{PV}(t) + P_{WT}(t) + P_{dis}(t) - P_{cha}(t) - P_{load}(t) + P_{DR}(t) + P_{buy}(t) - P_{sell}(t) = 0 \tag{6}$$

Here, $P_{MT}(t)$ represents the output power of microturbines at time t; $P_{PV}(t)$ and $P_{WT}(t)$ denote the output powers of photovoltaic and wind turbines at time t, respectively; $P_{dis}(t)$ and $P_{cha}(t)$ represent the discharge and charge powers of energy storage systems at time t; $P_{load}(t)$ is the load demand at time t; $P_{DR}(t)$ is the load reduction due to demand response programs at time t; and $P_{buy}(t)$ and $P_{sell}(t)$ are the power purchased from and sold to the main grid at time t. Equation (6) ensures that the power balance is maintained at any time point, thus ensuring the stable operation of the system.

3.3.2 Distributed Generator (MT) Output Limits

Microturbines operate within a feasible power range constrained by their technical specifications:

$$\min P_{MT} \leq P_{MT}(t) \leq \max P_{MT} \tag{7}$$

2025 S.T. Yau High School Science Award 仅用于2025丘成桐中学科学奖论文公示

Here, $\min P_{MT}$ and $\max P_{MT}$ represent the minimum and maximum output powers of microturbines, respectively. This constraint ensures that microturbines operate within a safe and efficient range, avoiding equipment damage or performance degradation caused by exceeding their design limits.

3.3.3 Battery Charging/Discharging Limits

To ensure safe operation and prolong battery life, charging and discharging powers are restricted:

$$P_{cha}(t) \geq 0, \quad P_{dis}(t) \leq \max P_{ESS}(t) \tag{8}$$

Here, $P_{cha}(t)$ denotes the charging power at time t, $P_{dis}(t)$ denotes the discharging power at time t, and $\max P_{ESS}(t)$ represents the maximum charge/discharge power of the energy storage system at time t. These constraints ensure that batteries operate within a safe range, avoiding damage caused by overcharging or over-discharging.

3.3.4 State of Charge (SOC) Dynamics

The state of charge (SOC) of the battery evolves according to charging/discharging activities and associated efficiencies:

$$E_{ESS}(t+1) = E_{ESS}(t) + P_{cha}(t) \cdot \eta_{cha} \cdot \Delta t - \frac{P_{dis}(t)}{\eta_{dis}} \cdot \Delta t \tag{9}$$

Here, $E_{ESS}(t)$ represents the energy stored in the battery at time t; $P_{cha}(t)$ and $P_{dis}(t)$ denote the charging and discharging powers at time t, respectively; η_{cha} and η_{dis} represent the charging and discharging efficiencies, respectively; and Δt is the duration of the time step. Equation (9) describes the evolution of battery energy over time, ensuring that the battery charges and discharges within reasonable limits, thereby maintaining the long-term stable operation of the system.

3.3.5 SOC Limits

The state of charge (SOC) of the battery must remain within the manufacturer-recommended range to avoid damage. This can be expressed as:

$$\min \mathrm{SOC} \leq \mathrm{SOC}(t) = \frac{E_{ESS}(t)}{\max E_{ESS}} \leq \max \mathrm{SOC} \tag{10}$$

Here, $\min \mathrm{SOC}$ and $\max \mathrm{SOC}$ represent the minimum and maximum allowable values for the battery's SOC; $E_{ESS}(t)$ denotes the energy stored in the battery at time t; and $\max E_{ESS}$ represents the maximum energy storage capacity of the battery. By setting these limits, we ensure that the battery operates within a safe range, preventing damage caused by overcharging or deep discharging.

3.3.6 Grid Exchange Limits

The power exchanged with the main grid cannot exceed the physical capacity of the connection point to ensure electrical safety, prevent overloading, and avoid tripping of protective devices. Specifically, the power purchased from and sold to the main grid must satisfy the following conditions:

$$0 \leq P_{buy}(t), \quad P_{sell}(t) \leq \max P_{EX} \tag{11}$$

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐科学奖论文公示

Here, $P_{buy}(t)$ represents the power purchased from the main grid at time t; $P_{sell}(t)$ represents the power sold to the main grid at time t; and $\max P_{EX}$ is the maximum power exchange capacity of the connection point. These constraints ensure that the power exchange between the system and the main grid remains within a safe limit, avoiding electrical faults caused by exceeding the physical capacity.

3.3.7 Demand Response Limitation

In demand response programs, the amount of load reduction must be limited to avoid negatively impacting user satisfaction. Excessive curtailment of power consumption can cause discomfort or inconvenience to users, thus the demand response constraint ensures that the reduction is kept within an acceptable range:

$$0 \leq P = \frac{P_{red}(t)}{P_{load}(t)} \leq \max P \tag{12}$$

Here, $P_{red}(t)$ denotes the reduced load at time t; $P_{load}(t)$ represents the total load demand at time t; and $\max P$ is the maximum allowable load reduction ratio. By setting this constraint, we can meet the operational needs of the system while ensuring that the user experience is not significantly affected.

4 Markov Decision Process Framework for Microgrid Optimization

4.1 MDP Formulation

The microgrid energy management problem is formally modeled as a Markov Decision Process (MDP), defined by the 4-tuple:

$$\text{MDP} = (S, A, P, r) \tag{13}$$

In this formulation, S represents the state space which encompasses all possible states of the system; A denotes the action space consisting of all feasible actions that can be taken at any given state; P is the state transition probability that defines the likelihood of transitioning from one state to another given a specific action; and r stands for the reward function which quantifies the immediate benefit received after taking an action in a particular state.

4.1.1 State Space (S)

The state vector S_t integrates external environmental inputs and internal operational status:

$$S_t = \{P_{WT}(t), P_{PV}(t), P_{load}(t), C_{buy}(t), C_{sell}(t), SOC_c\} \tag{14}$$

Here, $P_{WT}(t)$ represents the output power of wind turbines at time t; $P_{PV}(t)$ denotes the output power of photovoltaic systems at time t; $P_{load}(t)$ represents the load demand at time t; $C_{buy}(t)$ and $C_{sell}(t)$ denote the cost of purchasing power from the main grid and the revenue from selling power to the main grid at time t, respectively; and SOC_c represents the state of charge of the energy storage system. These variables collectively form a comprehensive description of the microgrid's operational state at any given time point.

2025 S-T-Yau High School Science Award 应用于2025丘成桐中学科学奖论文公示

4.1.2 Action Space (A)

The action vector is defined as:

$$A_t = (P_{MT}(t), P_{ESS}(t), P_{EX}(t)) \tag{15}$$

Here, $P_{ESS}(t)$ replaces the separate charge $P_{cha}(t)$ and discharge $P_{dis}(t)$ variables, while $P_{EX}(t)$ replaces the separate grid purchase $P_{buy}(t)$ and sale $P_{sell}(t)$. By using a single continuous variable for ESS and grid exchange, we ensure that the energy storage system cannot charge and discharge simultaneously, and the grid cannot buy and sell power at the same time.

4.1.3 Reward Function (r)

The reward function r_t is designed to guide the scheduling agent toward operational strategies that minimize cost while ensuring compliance with system constraints. It comprises two components: the dispatch cost $C_{cost}^{(t)}$ and the penalty term $C_{pen}^{(t)}$:

$$r_t = -(C_{cost}^{(t)} + C_{pen}^{(t)}) \tag{16}$$

The dispatch cost represents the direct monetary expenditure of the operator at time t, expressed as:

$$C_{cost}^{(t)} = C_{buy}^{(t)} - C_{sell}^{(t)} + C_{MT}^{(t)} \tag{17}$$

A lower $C_{cost}^{(t)}$ corresponds to reduced operating expenses, thereby increasing the reward. The penalty term discourages violations of operational constraints that could compromise system safety or reliability. Two categories of constraints are considered:

1. Directly avoidable constraints, which can be fully prevented by limiting the action space, thus requiring no penalty.
2. Indirect constraints, which cannot be eliminated through action limits alone and therefore require explicit penalty terms.

The penalty term is formulated as:

$$C_{pen}(t) = V_1 f_{SOC} + V_2 f_{EX} \tag{18}$$

Here, f_{SOC} quantifies the extent to which the battery state-of-charge (SOC) exceeds the safe range, and f_{EX} measures the excess in power exchange with the main grid beyond its physical limits. Each violation magnitude f_k is calculated as:

$$f_k = \ln\left(\frac{|X_k - \max X_k| + |X_k - \min X_k|}{\max X_k - \min X_k}\right) \tag{19}$$

where $\max X_k$ and $\min X_k$ are the allowable upper and lower bounds, and V_1, V_2 are the corresponding penalty coefficients.

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

4.2 Objectives

4.2.1 Cumulative Reward

In the Markov Decision Process (MDP) formulation, the objective is to maximize the cumulative discounted reward R_t over the decision horizon. The cumulative reward starting from time step t is defined as:

$$R_t = r_t + \gamma \cdot r_{t+1} + \gamma^2 \cdot r_{t+2} + ... + \gamma^{T-t} \cdot r_T \tag{20}$$

where r_t is the immediate reward obtained at time t, and γ is the discount factor with $0 \leq \gamma \leq 1$. In this study, we set $\gamma = 0.9$ to balance the relative importance of future rewards against immediate rewards.

4.2.2 Value Functions

The value function provides a measure for evaluating the quality of states and the benefits of actions under a given policy π.

1.State-value Function

The state-value function is defined as:

$$V_\pi(s) = E[R_t \mid S_t = s, \pi] \tag{21}$$

This equation represents the expected cumulative reward starting from state s and following policy π. It measures the average return that can be obtained by taking a series of actions according to policy π from a specific state.

2.Action-value Function

The action-value function is defined as:

$$Q_\pi(s, a) = E[R_t \mid S_t = s, A_t = a, \pi] \tag{22}$$

This equation represents the expected cumulative reward when taking action a in state s and thereafter following policy π. It measures the immediate and future returns after selecting a particular action in a specific state.

3.Optimal Policy

The optimal policy is determined by:

$$\pi(s) = \arg\max_a Q(s, a) \tag{23}$$

This equation indicates choosing the action a that maximizes the action-state value function for the current state s. The optimal policy aims to maximize long-term cumulative rewards by selecting the best action at each state.

5 Solution Methodology

This section presents the solution approach based on an improved Deep Q-Network (DQN) framework, designed to address overestimation issues and enhance learning stability in the micro-grid scheduling problem formulated in Section 3.

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

5.1 Neural Network Architecture

The Q-network is implemented as a neural network with the following structure:

- **Input Layer**: Processes the state vector representing real-time operating conditions, including electricity prices, state of charge, and load demand.
- **Hidden Layers**: Multiple fully-connected layers utilizing Rectified Linear Unit (ReLU) activations to extract high-dimensional features and model complex state relationships.
- **Output Layer**: Generates Q-values for all possible actions, employing a hyperbolic tangent (Tanh) activation function to constrain outputs within a stable numerical range.

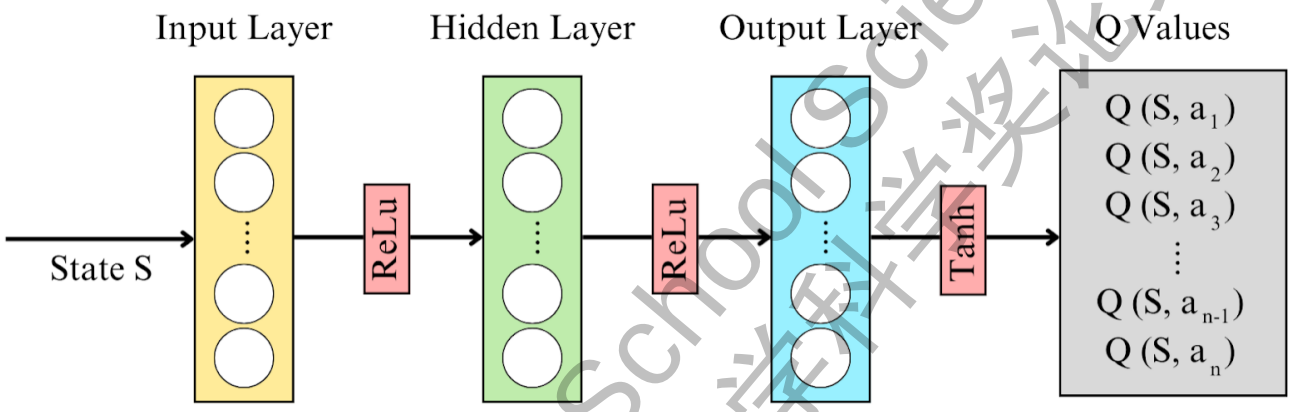

Figure 1: Neural Network Architecture

The input layer receives the state vector S, which contains information about the current environment, such as electricity prices, state of charge, and load demand. This layer enables the network to understand the current system's operational status and convert it into a form that can be processed by subsequent layers. The hidden layers consist of multiple fully connected layers, each using ReLU activation functions. The ReLU activation helps the network learn complex nonlinear relationships, thereby better capturing interactions between states. By stacking multiple layers, the network can gradually extract higher-level abstract features, which are crucial for accurately evaluating the value of different actions. The output layer is responsible for generating Q-values for all possible actions. To ensure the stability and reasonableness of the outputs, the Tanh activation function is employed here. The Tanh function constrains the output values between -1 and 1, helping to avoid numerical instability issues and making the network's learning process smoother and more reliable. This architecture captures nonlinear state-action dependencies while maintaining output stability, enabling optimal decision-making in dynamic grid environments.

5.2 Double Q-Network Design

To mitigate instability during training, a Double DQN architecture is adopted, consisting of two identical Q-networks:

1. Evaluation Q-network: interacts with the environment to select the optimal action based on the current state.

2025 S-Y Yau High School Science Award
仅用于2025年成功中学科学论文公示

2. Target Q-network: provides stable target Q-values for parameter updates.

The parameters of the target network are periodically updated from the evaluation network, preventing rapid target fluctuations and improving training stability.

5.3 Loss Function

The network parameters are optimized by minimizing the mean squared error between the predicted Q-values and target Q-values.

$$L(\theta) = E[(q_t - Q(s_t, a_t \mid \theta))^2] \tag{24}$$

Where: q_t is the target Q-value computed by the target network:

$$q_t = r_t + \gamma \max_{a_{t+1}} \overline{Q}(s_{t+1}, a_{t+1} \mid \overline{\theta}) \tag{25}$$

r_t is the immediate reward at time step t, representing the direct gain obtained from executing action a_t in state s. γ is the discount factor reflecting the depreciation rate of future rewards relative to immediate rewards. $\max_{a_{t+1}} \overline{Q}(s_{t+1}, a_{t+1} \mid \overline{\theta})$ is the maximum Q-value of the next state, given by the target network, representing the best possible future return.

5.4 Analysis of Overestimation in Deep Q-Networks

In conventional Deep Q-Networks (DQN), the target Q-value is computed using the maximum estimated Q-value over all possible actions in the next state:

$$q_t = r_t + \gamma \max_{a} Q(s_{t+1}, a) \tag{26}$$

However, this maximization operation introduces a positive bias in value estimation, leading to systematic overestimation of action values. This phenomenon was first identified by Thrun and Schwartz [11] and has since been extensively validated in reinforcement learning literature. Specifically, when the Q-function contains estimation errors—due to function approximation, limited data, or stochastic transitions—the term $\max_a Q(s_{t+1}, a)$ captures not only the true optimal action value but also amplifies noisy or high-variance estimates.

The overestimation problem degrades sample efficiency and may lead the agent to converge to a suboptimal policy. In dynamic energy systems characterized by fluctuating demand and intermittent renewable sources, inaccurate value assessments can trigger unnecessary control actions—such as excessive grid power purchases during peak pricing or inefficient cycling of energy storage units—thereby increasing operational costs and accelerating component degradation.

5.5 Proposed Framework: PER-Dueling-DDQN

To address the aforementioned limitations, this study proposes an integrated deep reinforcement learning architecture that combines multiple advanced techniques: Prioritized Experience Replay, Dueling Network streams, and Double DQN mechanisms, collectively referred to as PER-Dueling-DDQN. This framework improves learning performance through three complementary dimensions: network architecture design, target value computation, and experience utilization efficiency.

2025 S.T. Yau High School Science Award 仅用于2025丘成桐中学科学奖论文公示

5.5.1 Dueling Network Architecture

The proposed dueling network architecture decomposes the standard Q-value head into two separate streams: a state-value stream $V(s)$ and an advantage stream $A(s, a)$. This structural disentanglement enables the network to learn more nuanced representations of the environment's underlying dynamics. The state-value function $V(s)$ estimates the expected return from state s regardless of the selected action, reflecting the intrinsic desirability of being in that state. In contrast, the advantage function $A(s, a)$ quantifies how much better or worse a particular action a is compared to the average action in state s.

These components are combined to form the overall Q-value estimate:

$$Q(s, a) = V(s) + A(s, a) - \frac{1}{|A|} \sum_{a'} A(s, a') \tag{27}$$

The subtraction of the mean advantage term ensures identifiability of the decomposition, preventing ambiguity between $V(s)$ and $A(s, a)$. Without such normalization, multiple combinations of $V(s)$ and $A(s, a)$ could yield identical Q-values, destabilizing training and impairing generalization. Empirical studies [12] have demonstrated that dueling networks achieve faster convergence

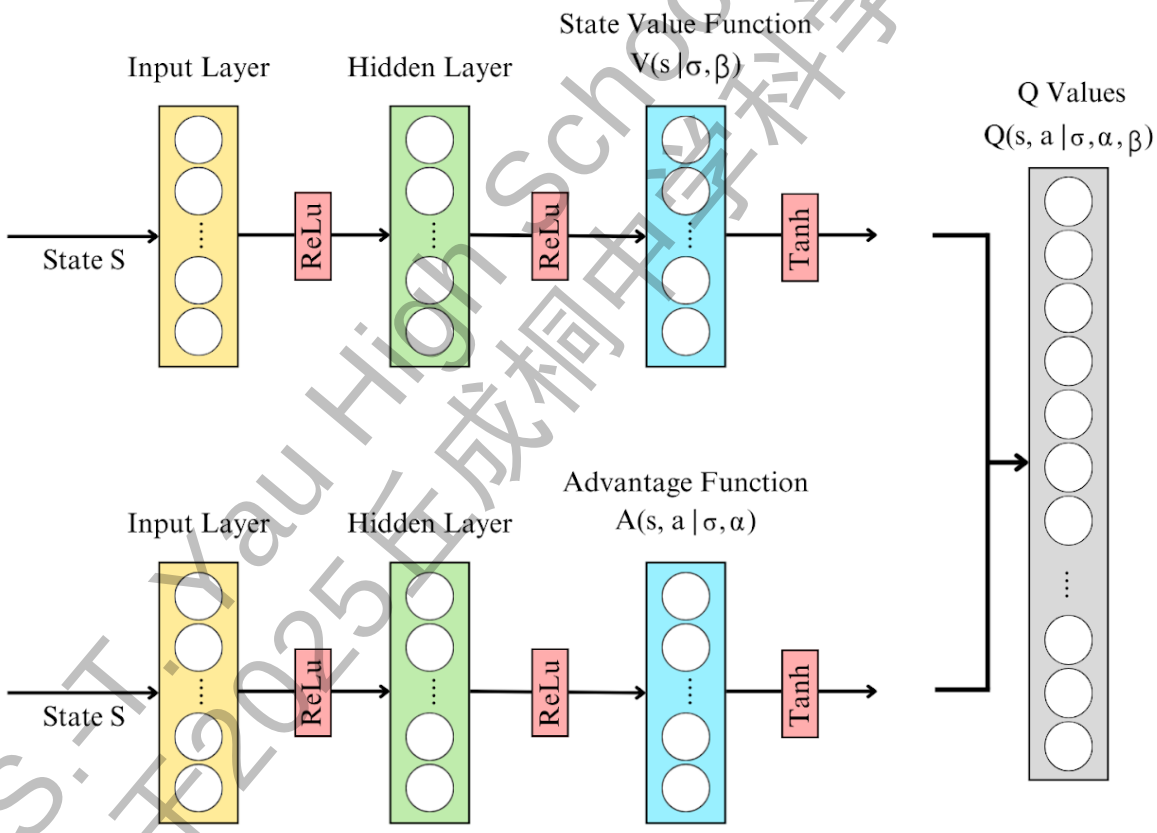

Figure 2: Architecture of PER-Dueling-DDQN

and improved policy quality compared to standard DQN, particularly in environments with sparse rewards and complex state dependencies. By sharing representation across value and advantage streams, the model reduces redundancy and enhances parameter efficiency, enabling more robust learning from limited interaction data.

In this architecture, the Q-value is obtained by combining two distinct network outputs:

2025 S.-T. Yau High School Science Award
仅用于2025年成桐中学科学奖公示

- State-value network $V(s \mid \sigma, \beta)$: a one-dimensional output that evaluates the overall value of being in a given state s.
- Action-advantage network $A(s, a \mid \sigma, \alpha)$: a $|A|$ dimensional output that evaluates the relative advantage of each action a in state s.

The final Q-value is expressed as:

$$Q(s, a \mid \sigma, \alpha, \beta) = V(s \mid \sigma, \beta) + \left[A(s, a \mid \sigma, \alpha) - \frac{1}{|A|} \sum_{a' \in A} A(s, a' \mid \sigma, \alpha) \right] \tag{28}$$

Where σ denotes the shared parameters of the two deep neural networks (DNNs), while α and β denote the independent parameters of the advantage and value networks, respectively. Thus, the full parameter set of the D3QN is given by $\theta = \{\sigma, \alpha, \beta\}$. Subtracting the mean advantage ensures that the advantage function has zero mean across all actions, thereby removing redundant degrees of freedom and improving stability. Consequently, the dueling structure reduces noise and instability during iterative updates. In summary, the double network design (decoupling action selection from target Q-value estimation) and the dueling structure (improving Q-value decomposition) jointly enhance the accuracy and robustness of Q-learning compared to the conventional DQN.

5.5.2 Prioritized Experience Replay (PER)

Instead of sampling experiences uniformly from the replay buffer, prioritized experience replay selects transitions with larger temporal-difference errors more frequently. This prioritization ensures that the agent focuses on high-error (i.e., more informative) experiences, accelerating convergence and improving performance.

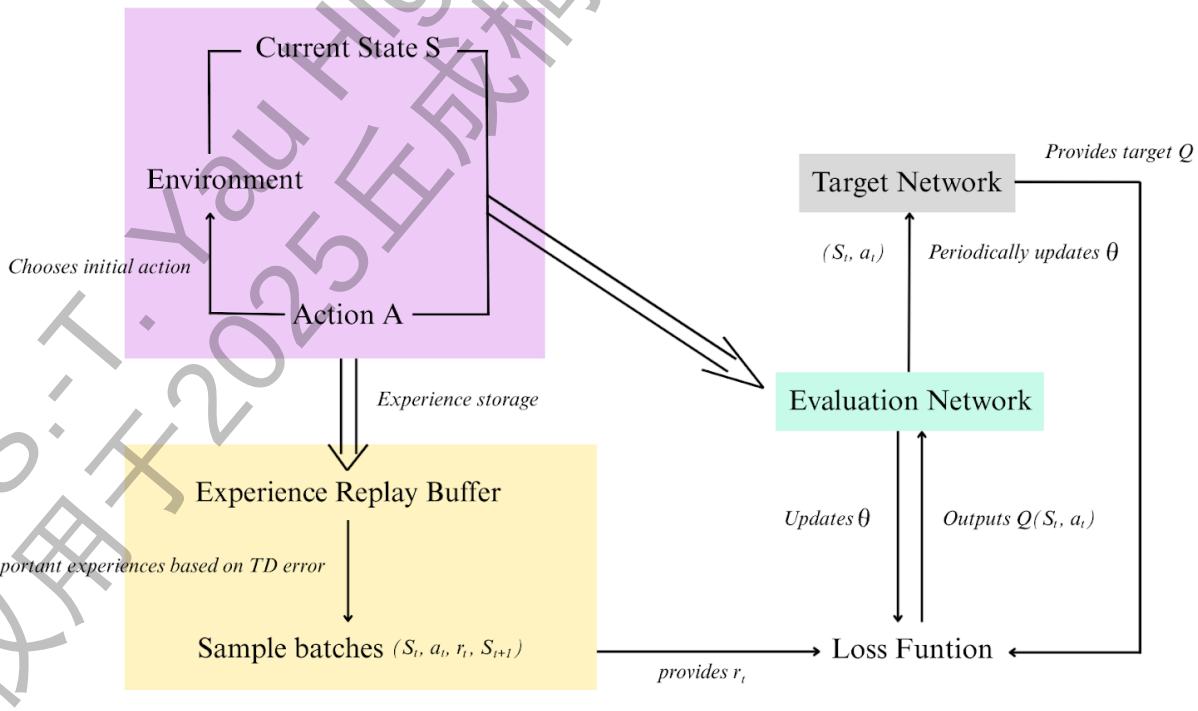

Figure 3: Prioritized Experience Replay (PER)

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖公示

5.6 Training Process

The training objective is to minimize the total operating cost of the microgrid through continuous interaction between the agent and the environment. The agent learns to predict the action-value function $Q(s, a)$ by balancing exploration and exploitation until the Q-values converge. The training procedure consists of the following five steps:

Step 1: Initialization

The Q-network is initialized with weights. The output layer corresponds to the number of discretized actions. The target value is defined as:

$$Q_{target} = r + \gamma \cdot \max Q(s', a') \tag{29}$$

Step 2: Action Discretization

Since the original action space is continuous (e.g., MT output, ESS charging/discharging, DR response), it is discretized to fit the Q-learning framework:

1. Define the action range for each dimension k: $[min\ x_k, max\ x_k]$.
2. Choose a discretization interval Δx_k.
3. Compute the number of discrete actions:

$$X_k = \frac{max\ x_k - min\ x_k}{\Delta x_k} + 1 \tag{30}$$

4. Generate discrete actions:

$$X_k = min\ x_k + \Delta x_k \cdot i \quad (i = 0, 1, ..., x_k - 1) \tag{31}$$

Step 3: Action Selection

The agent selects an action a_t using an ε-greedy strategy:

$$a_t = \begin{cases} a_{random} & \text{(exploration)} \quad \text{with probability } \varepsilon \\ \arg\max_a Q(S_t, a; \theta) & \text{(exploitation)} \quad \text{with probability } 1 - \varepsilon \end{cases}$$

This balances exploration (searching wider action space) and exploitation (using the best-known action). The value of ε decays over training to gradually shift from exploration to exploitation.

Step 4: Environment Interaction

The chosen action is executed in the environment, leading to state transition.

$$C_{total} = C_{MT} + C_{ESS} + C_{DR}$$

Experiences are stored in the replay buffer, and prioritized replay is applied to improve sample efficiency.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

Step 5: Parameter Update

The Q-network parameters θ are updated via SGD:

1. **Mini-Batch Sampling**

   Sample a batch $\{S_t, a_t, r_t, S_{t+1}\}$ from the replay buffer.

2. **Compute target Q-value**

   The discounted total future reward that can be obtained after executing the action is calculated by:

$$Q_{target} = r_t + \gamma \cdot \max_a Q(S_{t+1}, a'; \overline{\theta})$$

3. **Compute current predicted Q-value**

$$Q_{pred} = Q(S_t, a_t; \theta)$$

4. **Compute the loss function and minimize the mean squared error (MSE):**

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} (Q_{target,i} - Q_{pred,i})^2 \tag{32}$$

Where N is the batch size.

5. **Update parameters:**

$$\theta = \theta - \eta \cdot \nabla_\theta L(\theta) \tag{33}$$

6. **Update the target network.**

6 Coordinated Optimal Scheduling Based on Stackelberg Game and DDPG Algorithm

6.1 System Architecture

The considered distribution network architecture comprises a single Distribution System Operator (DSO) managing a radial feeder with multiple interconnected microgrids (MGs). Bi-directional power and communication flows exist between the DSO and each MG. The DSO can purchase electricity from the main grid and exchange power with MGs based on operational needs.

In this system, the DSO acts as the leader, guiding the behavior of MGs through Stackelberg game strategies to achieve global optimal scheduling. Each MG serves as a follower, adjusting its operational strategy according to the DSO's decisions to reach local optimality. Furthermore, the Deep Deterministic Policy Gradient (DDPG) algorithm is introduced to optimize the interaction process between the DSO and MGs, enhancing the overall efficiency and stability of the system.

Figure 4 illustrates the schematic diagram of the system architecture, where the DSO interacts with multiple MGs via energy flow and information flow. Each MG includes distributed energy resources such as photovoltaic (PV), wind turbines (WT), energy storage systems (ESS), and micro-turbines (MT). By coordinating the operation of these devices, the DSO effectively manages the entire distribution network.

2025 S-T Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

Algorithm 1 Per Dueling DDQN (D3QN)

Randomly initialize the evaluation network parameters θ
Initialize target network parameters $\theta \rightarrow \underline{\theta}$
for $(episode = 1) : E$ **do**
    Reinitialize the environment state
    for $(time\ step\ t = 0) : T$ **do**
        Obtain environment status
        Choose action a_t based on strategy π
        Calculate the immediate reward r_t and obtain the new environment state
        Store experience in the experience replay buffer
        if the experience replay pool experience ¿ number minibatch sample capacity (Minibatch Size) **then**
            Prioritize experience replay sampling from the experience replay pool
            if $t = T$ **then**
                $q_t = r_t$
            else
                $q_t = r_t + \gamma \cdot \max Q(S_{t+1}, a_{t+1})/\underline{Q}$
            end if
            Calculate the loss function $L(\theta)$
            Update the current network's parameters θ
            Update network parameters every C rounds
        end if
    end for
end for

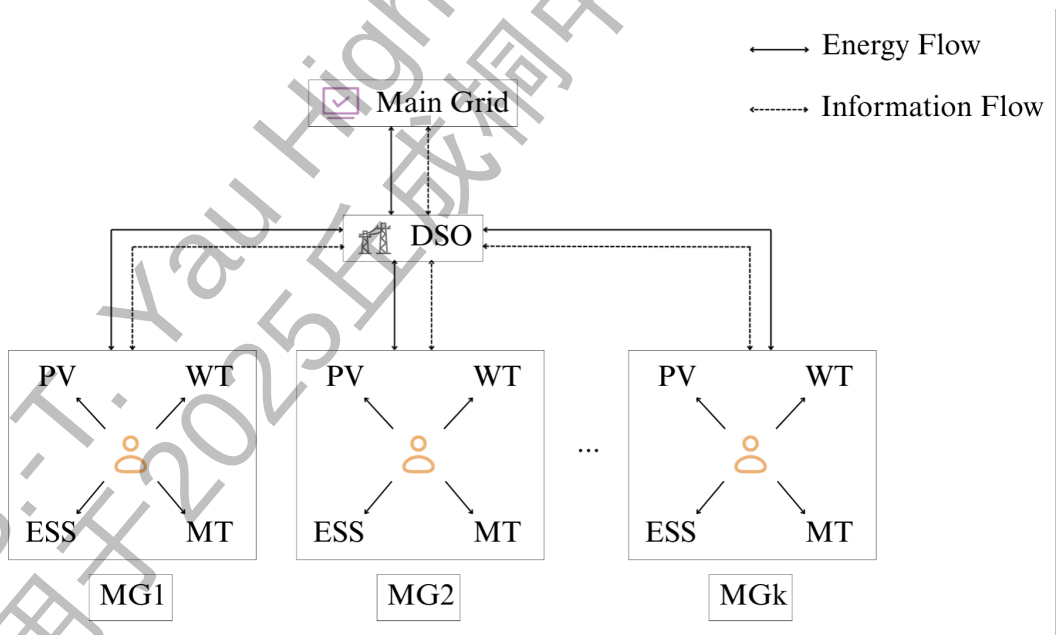

Figure 4: System Architecture

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

6.2 Upper-Level Optimization Model

6.2.1 Objective Function

The objective function aims to minimize the total operating cost for the Distribution System Operator (DSO) at time t, denoted as $C_{DSO,t}$. This cost comprises three key components: the cost of purchasing electricity from the wholesale market, the net cost of transactions with microgrids, and the cost of energy losses in the distribution lines.

$$\min C_{DSO,t} = C_{DSO,1,t} + C_{DSO,2,t} + C_{DSO,3,t} \tag{34}$$

Where:

- $C_{DSO,1,t}$ is the cost of purchasing power from the upper-level grid.
- $C_{DSO,2,t}$ is the net cost of energy transactions with all connected microgrids.
- $C_{DSO,3,t}$ is the cost of active power losses in the distribution network.

1) Cost of Power Purchased from Main Grid:

This cost is determined by the product of the electricity price $C_{maingrid,t}$ set by the wholesale market, the purchased power $P_{maingrid,t}$, and the time interval Δt.

$$C_{DSO,1,t} = C_{maingrid,t} \cdot P_{maingrid,t} \cdot \Delta t \tag{35}$$

By optimizing these three components, the DSO can effectively reduce its operational costs and enhance the economic efficiency of the entire distribution system. Moreover, rational planning and management of energy transactions with microgrids not only decrease reliance on the main grid but also promote the utilization of renewable energy sources, contributing to sustainable development goals.

2) Net Cost of Transactions with Microgrids:

This represents the net monetary flow between the DSO and N microgrids. The DSO incurs a cost when it purchases power from a microgrid i at the price $C_{MGs,i,t}$. Conversely, it generates revenue when it sells power to a microgrid i at the price $C_{MGb,i,t}$. The net cost is the sum of all purchases minus the sum of all sales across all microgrids.

$$C_{DSO,2,t} = \sum_{i=1}^{N} (C_{MGs,i,t} \cdot P_{MGs,i,t} - C_{MGb,i,t} \cdot P_{MGb,i,t}) \cdot \Delta t \tag{36}$$

By optimizing transactions with microgrids, the DSO can more effectively manage its energy supply and demand balance, reduce unnecessary cost expenditures, and enhance overall economic efficiency. Additionally, rational pricing strategies can encourage active participation from microgrids, collectively promoting the sustainable development of the distribution system.

3) Cost of Network Losses:

The financial impact of active power losses $P_{loss,j,t}$ across all N branches in the distribution network is calculated.

$$C_{DSO,3,t} = C_{maingrid,t} \cdot \left(\sum_{j=1}^{N} P_{loss,j,t} \right) \cdot \Delta t \tag{37}$$

Network loss costs reflect energy losses during power transmission, which have a direct impact on the DSO's operational costs. By adopting advanced technologies and management measures, network losses can be effectively reduced, improving power transmission efficiency. This further lowers operational costs and enhances system performance.

2025 S-T Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

6.2.2 Power Flow and Network Constraints

(1) Voltage Equation

$$V_{j+1,t}^2 = V_{j,t}^2 - 2(r_j P_{j,t} + x_j Q_{j,t}) + \frac{(r_j^2 + x_j^2)(P_{j,t}^2 + Q_{j,t}^2)}{V_{j,t}^2} \tag{38}$$

This equation demonstrates how the voltage magnitude changes from node j to node $j+1$ along branch j, taking into account the impact of branch resistance r_j, reactance x_j, and the injected active power $P_{j,t}$ and reactive power $Q_{j,t}$ at node j. By analyzing the voltage equation, a deeper understanding of the voltage distribution patterns in power systems can be gained, providing theoretical support for optimizing grid operations and improving power supply quality.

(2) Active Power Equation

$$P_{j+1,t} = P_{j,t} - r_j \frac{P_{j,t}^2 + Q_{j,t}^2}{V_{j,t}^2} - p_{j+1,t} \tag{39}$$

This equation reflects the active power balance along branch j. Studying the active power equation helps to understand the characteristics of energy transmission in power systems, which is crucial for rational planning of power dispatching and ensuring the safe and stable operation of the grid.

(3) Reactive Power Equation

$$Q_{j+1,t} = Q_{j,t} - x_j \frac{P_{j,t}^2 + Q_{j,t}^2}{V_{j,t}^2} q_{j+1,t} \tag{40}$$

This equation describes the reactive power balance state. The role of reactive power in power systems cannot be overlooked as it directly affects voltage levels and system stability. In-depth research on the reactive power equation can effectively enhance the operational efficiency and reliability of power systems, laying a solid foundation for achieving the goals of smart grids.

6.3 Lower-Level Optimization Model

6.3.1 Objective Function

The objective of each microgrid i is to minimize its total operational cost at time t:

$$\min C_{MG,i,t} = C_{MG,1,i,t} + C_{MG,2,i,t} + C_{MG,3,i,t} \tag{41}$$

Where the cost components are defined as follows:

(1) Energy Trading Cost with DSO

$$C_{MG,1,i,t} = (C_{MGs,i,t} \cdot P_{MGs,i,t} - C_{MGb,i,t} \cdot P_{MGb,i,t}) \cdot \Delta t \tag{42}$$

Where $C_{MGs,i,t}$ and $C_{MGb,i,t}$ are electricity prices, $P_{MGs,i,t}$ and $P_{MGb,i,t}$ are electricity quantities for different transaction directions, and Δt is the time interval. By optimizing energy trading with the DSO, the microgrid can effectively reduce operational costs and enhance economic benefits.

(2) Micro-turbine Operation Cost

$$C_{MG,2,i,t} = (a + b \cdot P_{MT,i,t}) \cdot \Delta t \tag{43}$$

Where a and b are cost coefficients, and $P_{MT,i,t}$ is the output power of the micro-turbine. Proper configuration of the micro-turbine's operating parameters can significantly reduce fuel consumption and maintenance costs, thereby lowering overall operational costs.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

(3) Energy Storage Operation Cost

$$C_{MG,3,i,t} = C_{ESS} \cdot (P_{cha,i,t} \cdot \eta_{cha} + \frac{P_{dis,i,t}}{\eta_{dis}}) \cdot \Delta t \tag{44}$$

Where C_{ESS} is the energy storage cost coefficient, $P_{cha,i,t}$ and $P_{dis,i,t}$ are charging and discharging powers, and η_{cha} and η_{dis} are charging and discharging efficiencies. Through optimized management of the energy storage system, efficient utilization of electricity can be achieved, further reducing the operational costs of the microgrid.

6.3.2 Decision Variables

- $P_{MT,i,t}$: Output power of the micro-turbine.
- $P_{cha,i,t}, P_{dis,i,t}$: Charging and discharging powers of the energy storage system.
- $P_{MGs,i,t}, P_{MGb,i,t}$: Power exchange with the DSO.

6.3.3 Constraints

(1) Power Balance Constraint

The total power from the micro-turbine, photovoltaic (PV), wind turbine (WT), and energy storage charging/discharging should balance the load and transaction power. This constraint ensures the energy supply-demand balance within the microgrid, which is fundamental for maintaining stable operation.

$$P_{MT,i,t} + P_{PV,i,t} + P_{WT,i,t} - P_{cha,i,t} + P_{dis,i,t} = P_{load,i,t} + P_{MGs,i,t} - P_{MGb,i,t} \tag{45}$$

By precisely controlling the power output of each generation unit and storage device, effective responses to load fluctuations and changes in the external power market can be achieved, enhancing the self-sufficiency and economic efficiency of the microgrid.

(2) Micro-turbine Operation Constraint

The output power of the micro-turbine should be within the minimum and maximum limits to ensure safe and stable operation.

$$\min P_{MT} \leq P_{MT,i,t} \leq \max P_{MT} \tag{46}$$

Reasonably setting the operating range of the micro-turbine can prevent overloading or inefficient operation, extending the equipment's lifespan and reducing maintenance costs. **(3) Energy Storage Power Constraint**

The charging and discharging powers of the energy storage system should be within the minimum and maximum limits to ensure safe operation and prevent overcharging or over-discharging.

$$\min P_{ESS} \leq P_{cha,i,t} \leq \max P_{ESS}, \quad 0 \leq P_{dis,i,t} \leq \max P_{ESS} \tag{47}$$

Properly setting the power range of the energy storage system can effectively extend the battery life, improve overall efficiency, and enhance system reliability.

(4) Energy Storage SOC Constraint

The state of charge (SOC) of the energy storage system should be maintained within the minimum and maximum limits to ensure normal operation and prevent excessive charging or discharging.

$$\min SOC \leq \frac{E_{ESS,i,t}}{\max E_{ESS}} \leq \max SOC \tag{48}$$

2025 S.-T. Yau High School Science Award 仅用于2025丘成桐中学科学奖论文公示

The SOC update formula is described as:

$$E_{ESS,i,t+1} = E_{ESS,i,t} + \eta_{cha} \cdot P_{cha,i,t} \cdot \Delta t - \frac{P_{dis,i,t}}{\eta_{dis}} \cdot \Delta t \tag{49}$$

By precisely controlling the SOC, the use of the energy storage system can be optimized, improving the energy management efficiency of the microgrid.

6.4 Network Constraints

(1) Branch Power Loss

The active power loss of branch j at time t is:

$$P_{loss,j,t} = r_{j,t} \cdot \frac{P_{j,t}^2 + Q_{j,t}^2}{V_{j,t}^2} \tag{50}$$

(2) Node Voltage Constraint

The voltage at node j at time t should be within the allowable range:

$$\min V \leq V_{j,t} \leq \max V \tag{51}$$

(3) Branch Power Constraint

The active power on branch j at time t satisfies:

$$0 \leq P_{j,t} \leq \max P_{line,j} \tag{52}$$

Where $\max P_{line,j}$ is the maximum transmission power of the distribution network feeder. This constraint ensures the safe and stable operation of the grid, preventing overloading and potential faults.

(4) Electricity Price Constraint

The electricity price for transactions between the distribution network and microgrid i is bounded:

$$\min C_{MG,i} \leq C_{MGb,i,t} \leq C_{MGs,i,t} \leq \max C_{MG,i} \tag{53}$$

Where $\min C_{MG,i}$ and $\max C_{MG,i}$ are the minimum and maximum electricity prices for transactions between the network and microgrid i. Reasonably setting the price range can promote fair trading and protect the interests of all parties involved.

6.5 Stackelberg Game Framework

The interaction between the Distribution System Operator (DSO) and multiple microgrids (MGs) is formulated as a Stackelberg game, where the DSO acts as the leader and the MGs are the followers. The DSO determines energy trading prices and dispatch decisions, while each MG optimizes its local operation strategy based on the DSO's decisions.

6.5.1 Sets of Participants

- φ_{DSO}: Sets of DSOs
- φ_{MG}: Sets of microgrids indexed by $i \in \varphi_{MG}$

2025 S-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

6.5.2 Strategy Sets

The DSO decides the buying and selling electricity prices at time t:

$$X_{DSO,t} = f(C_{MGs,i,t}, C_{MGb,i,t} | i \in \varphi_{MG}) \tag{54}$$

Where: $C_{MGs,i,t}$ represents the electricity selling price from the Distribution System Operator (DSO) to microgrid i, while $C_{MGb,i,t}$ denotes the electricity buying price from microgrid i to the DSO. The time step index t ranges from 1 to T, tracking transactions at different time points.

Each microgrid adjusts its power scheduling decisions based on these price signals to minimize operational costs and maximize utility. Specifically, the strategy set $X_{MG,i,t}$ for microgrid i at time t includes several components:

$$X_{MG,i,t} = \{P_{MT,i,t}, P_{cha,i,t}, P_{dis,i,t}, P_{MGs,i,t}, P_{MGb,i,t}\} \tag{55}$$

Where:

- $P_{MT,i,t}$: Internal generation within the microgrid.
- $P_{cha,i,t}$: Charging power.
- $P_{dis,i,t}$: Discharging power.
- $P_{MGs,i,t}$: Electricity sold to the DSO.
- $P_{MGb,i,t}$: Electricity purchased from the DSO.

6.5.3 Stackelberg Game Model Construction

We construct a Stackelberg game model with centralized coordination and decentralized control.

(1) Dynamic Process

The Distribution System Operator (DSO) formulates electricity prices strategy, while microgrids (MGs) adjust their operation strategies. Through iterative convergence, the system reaches equilibrium. This dynamic process ensures that the system can achieve optimal configuration under constantly changing market conditions, thereby realizing effective resource allocation and utilization.

(2) Key Components

The game model aggregates core elements to form a complete decision-making system.

$$\Psi = \{\phi_{DSO}, \phi_{MG}, X_{DSO,t}, \{X_{MG,i,t} | i \in \phi_{MG}\}, U_{DSO,t}, \{U_{MG,i,t} | i \in \phi_{MG}\}\} \tag{56}$$

Where ϕ_{DSO} and ϕ_{MG} are participant sets (defined in section 6.5.1), $X_{DSO,t}$ and $X_{MG,i,t}$ are strategy sets (defined in section 6.5.2), and $U_{DSO,t}$ and $U_{MG,i,t}$ are utility functions.

(3) Coordinated Mechanism

Centralized Coordination: The DSO aggregates information from all microgrids to formulate price strategies that match the overall grid status. This centralized coordination mechanism ensures global optimization, enabling each microgrid to make decisions within a unified framework, thus enhancing the efficiency and stability of the entire system.

Decentralized Control: Each microgrid independently adjusts its operation strategy based on the price information provided by the DSO, aiming to minimize operational costs and maximize benefits. This decentralized control mechanism grants microgrids a certain degree of autonomy, allowing them to flexibly respond to local demands and resource changes.

2025 S. T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

By combining centralized coordination and decentralized control, the Stackelberg game model achieves a balance between global optimization and local flexibility, providing strong support for the efficient operation of power systems.

(4) Utility Function

In the bi-level optimal dispatch model of the distribution network and microgrids, the utility functions represent the economic interests of different participants. Since each participant aims to minimize operational costs, maximizing utility is mathematically equivalent to minimizing these costs. Therefore, the utility functions are expressed as the negative of the respective cost functions.

i) Utility Function of DSO ($U_{DSO,t}$)

It represents the interests of the Distribution System Operator (DSO). The maximization of its utility is equivalent to the minimization of operation cost. The operation cost of the distribution network includes the cost of purchasing electricity from the main grid, network loss cost, etc. The expression is:

$$U_{DSO,t} = -F_{DSO,t} \tag{57}$$

where $F_{DSO,t}$ is the operation cost of the distribution network at time t.

ii) Utility Function of Microgrids ($U_{MG,i,t}$)

It represents the interests of the Microgrid (MG). The maximization of utility corresponds to the minimization of its own operation cost. The operation cost of the microgrid includes the fuel cost of the micro-turbine, the depreciation cost of energy storage, the transaction cost with the distribution network, etc. The expression is:

$$U_{MG,i,t} = -f_{MG,i,t} \tag{58}$$

where $f_{MG,i,t}$ is the operation cost of microgrid i at time t.

The interaction among multi-agents is promoted by the utility function, following the process of "distribution network formulates price strategy → microgrid responds with strategy → distribution network evaluates results". This mechanism ensures that while each participant pursues its own interests, it also contributes to the overall optimization and stable operation of the system.

iii) Stackelberg Equilibrium

In the context of hierarchical energy management between a Distribution System Operator (DSO) and multiple Microgrids (MGs), the strategic interaction is modeled as a Stackelberg game, where the DSO acts as the leader and the MGs as followers. The solution concept of this non-cooperative game is known as the *Stackelberg equilibrium*, defined as a strategy profile from which no participant can benefit by unilaterally altering its decision, given that the followers respond rationally to the leader's action.

At equilibrium, the DSO first commits to a pricing strategy, characterized by the electricity purchase price $C_{MGb,i,t}$ and sale price $C_{MGs,i,t}$ for each microgrid i at time t. Each MG, upon receiving these price signals, independently solves its local optimization problem to determine an optimal operational schedule $X^*_{MG,i,t}$—comprising controllable variables such as $P_{MT,i,t}$ (micro-turbine output), $P_{cha,i,t}$, $P_{dis,i,t}$ (energy storage power), and power exchange terms $P_{MGb,i,t}$, $P_{MGs,i,t}$. This decision maximizes its utility $U_{MG,i,t} = -f_{MG,i,t}$, where $f_{MG,i,t}$ denotes the operational cost, subject to physical and operational constraints.

Anticipating the rational responses of all MGs, the DSO formulates its pricing policy $X^*_{DSO,t}$ to maximize its own utility $U_{DSO,t} = -F_{DSO,t}$, where $F_{DSO,t}$ represents system-level costs including grid procurement and losses. This sequential decision process is mathematically captured by a bilevel optimization framework, formally expressed as:

2025 S.T. Yau High School Science Award 仅用于2025丘成桐中学科学奖论文公示

$$
\begin{aligned}
\max_{X_{DSO,t}} \quad & U_{DSO,t} = -F_{DSO,t}(X_{DSO,t}, \{X^*_{MG,i,t} \mid i \in \phi_{MG}\}) && \text{(59a)}\\
\text{s.t.} \quad & \text{Network constraints: } V_{\min} \le V_j \le V_{\max},\ \forall j \in \mathcal{N} && \text{(59b)}\\
& C_{MGb,i,t}, C_{MGs,i,t} \in [C_{\min}, C_{\max}],\ \forall i \in \phi_{MG} && \text{(59c)}\\
& \text{For each } i \in \phi_{MG},\ X^*_{MG,i,t} \text{ solves:} \\
& \quad \max_{X_{MG,i,t}} \quad U_{MG,i,t} = -f_{MG,i,t}(X_{MG,i,t}, X_{DSO,t}) \\
& \quad \text{s.t.} \quad P_{cha,i,t} \le P^{\max}_{cha},\ P_{dis,i,t} \le P^{\max}_{dis} && \text{(59d)}\\
& \qquad E_{i,t+1} = E_{i,t} + \eta_{cha} P_{cha,i,t} - \eta^{-1}_{dis} P_{dis,i,t} \\
& \qquad P^{\min}_{MT} \le P_{MT,i,t} \le P^{\max}_{MT}
\end{aligned}
$$

Here, the upper-level problem (1a)–(1c) corresponds to the DSO's strategic decision-making, while the lower-level problems (1d)–(1e) represent the decentralized responses of individual MGs, parameterized by the leader's announced prices. The Stackelberg equilibrium is attained when a fixed point $(X^*_{DSO,t}, \{X^*_{MG,i,t}\})$ is reached, such that both levels simultaneously satisfy optimality conditions.

6.6 Solution via Policy-Based Deep Reinforcement Learning

Given the computational complexity and non-convex nature of the bilevel optimization problem, an analytical solution is often intractable, especially in large-scale or uncertain environments. To address this challenge, we adopt a policy-based Deep Reinforcement Learning (DRL) framework, specifically the Deep Deterministic Policy Gradient (DDPG) algorithm, to approximate the optimal pricing policy for the DSO.

The interaction between the DSO and the MGs is formulated as a Markov Decision Process (MDP), where the DSO functions as the learning agent. The state space $\mathcal{S}$ encompasses system-wide observables, including load profiles, renewable generation forecasts, voltage levels, and historical price signals. The action space $\mathcal{A}$ is defined by the DSO's pricing vector $X_{DSO,t}$. The reward at each time step is designed as $r_t = U_{DSO,t}$, aligning the agent's objective with the system's economic performance.

Crucially, the MGs' responses are treated as part of the environment dynamics, either through simulation or embedded optimization modules. This allows the DRL agent to learn an implicit reaction function without explicit modeling of follower objectives, thereby reducing dependence on precise knowledge of internal MG cost structures. The DDPG algorithm, leveraging actor-critic architecture and experience replay, enables stable learning of continuous control policies in high-dimensional spaces, making it well-suited for real-time, adaptive energy management in distribution networks.

6.6.1 Markov Decision Process Construction

The Markov Decision Process (MDP) framework defines the state space, action space, and reward functions for both the Distribution System Operator (DSO) and microgrids.

(1) State Space

As the leader in the game, the DSO's state space reflects the market environment and the expected power trading information submitted by microgrids:

$$
S_{DSO,t} = \{P_{MG,1,t}, \ldots, P_{MG,k,t}, C_{maingrid,t}\} \tag{60}
$$

2025 S-T Yau High School Science Award 仅用于2025丘成桐中学科学奖论文公示

Where:

- $P_{MG,i,t}$ $(i = 1, 2, \ldots, k)$ represents the expected transaction power submitted by microgrid i at time t.
- $C_{maingrid,t}$ is the wholesale market transaction electricity price at time t.

As the follower in the game, the microgrid's state space must include the decision feedback from the distribution network and the operating status of internal equipment to support strategy response. Its expression is:

$$S_{MG,i,t} = \{C_{MGs,i,t}, C_{MGb,i,t}, P_{PV,i,t}, P_{WT,i,t}, P_{load,i,t}, SOC_{i,t}\} \quad i \in k \tag{61}$$

Where:

- $C_{MGs,i,t}$ and $C_{MGb,i,t}$ represent the selling and buying prices of microgrid i at time t, respectively.
- $P_{PV,i,t}$ and $P_{WT,i,t}$ represent the photovoltaic and wind turbine generation powers of microgrid i at time t, respectively.
- $P_{load,i,t}$ represents the load demand of microgrid i at time t.
- $SOC_{i,t}$ represents the state of charge of the energy storage system of microgrid i at time t.

By defining these state spaces, a comprehensive reflection of the market environment and operational states of both the DSO and microgrids at different time points can be achieved, providing a solid foundation for subsequent action selection and reward calculation.

(2) Action Space

The Distribution System Operator (DSO) determines the buying and selling prices for electricity trading with each microgrid:

$$A_{DSO,t} = \{C_{MGs,1,t}, \ldots, C_{MGs,k,t}, C_{MGb,1,t}, \ldots, C_{MGb,k,t}\} \tag{62}$$

Each microgrid decides internal equipment operation (e.g., microturbine generation, battery charging/discharging) and expected transaction power with the DSO:

$$A_{MG,i,t} = \{P_{MT,i,t}, P_{BSS,i,t}\} \tag{63}$$

(3) Reward Functions

The DSO aims to minimize operational costs while ensuring grid-safe operation. Its reward function incorporates both operational costs and penalty terms for constraint violations:

$$r_{DSO,t} = -(C_{DSO,t} + C_{DSO,pen,t}) \tag{64}$$

Where:

- $C_{DSO,t}$ represents the base operational cost of the network at time t.
- $C_{DSO,pen,t}$ is the penalty cost for violating operational constraints; it aggregates penalties for voltage violations and line overloads:

$$C_{DSO,pen,t} = w_1 f_v + w_2 f_{line} \tag{65}$$

w_1 and w_2 are weight coefficients for voltage and line overload violations, respectively. f_v and f_{line} are penalty functions triggered by voltage limits and line load exceeding capacity.

2025 S.-T. Yau High School Science Award 仅用于2025年丘成桐中学科学奖论文公示

For each microgrid, the reward function is expressed as:

$$r_{MG,i,t} = -(C_{MG,i,t} + C_{MG,pen,i,t}) \quad i \in k \tag{66}$$

Where:

- $C_{MG,i,t}$ is the base operational cost for violating microgrid-specific constraints.
- $C_{MG,pen,i,t}$ is the penalty cost:

$$C_{MG,pen,i,t} = v_1 f_{SOC,i} + v_2 f_{ex,i} \tag{67}$$

v_1 and v_2 are weight coefficients for SOC violation and trading excess. $f_{SOC,i}$ and $f_{ex,i}$ are penalty functions activated by SOC out of safe range and power exchange exceeding contracted limits.

6.6.2 Deep Deterministic Policy Gradient (DDPG) Algorithm

To solve the continuous decision-making problem in the bi-level Stackelberg game framework, this work adopts the Deep Deterministic Policy Gradient (DDPG) algorithm.

(1) Leader-Follower Interaction

As illustrated on the left side of the figure, the leader agent (DSO) determines its action $a_{t,DSO}$, which corresponds to the electricity trading prices $C_{MGs,i,t}$ and $C_{MGb,i,t}$. The follower agents (Microgrids, MGs) observe these price signals along with their local states — such as photovoltaic generation $P_{PV,i,t}$, wind turbine output $P_{WT,i,t}$, and load demand $P_{load,i,t}$ — and compute their optimal response actions $a_{MG,i,t}$.

For each time step:

1. The DSO sets price strategies $a_{t,DSO}$.
2. MGs optimize their internal scheduling strategies $a_{MG,i,t}$ accordingly.
3. MGs return their decisions and expected transaction power $P_{MGs,i,t}$ and $P_{MGb,i,t}$ to the DSO.
4. Both the DSO and MGs update their states $s_{DSO,t+1}$ and $s_{MG,t+1}$ and corresponding reward.

This iterative interaction enables the system to gradually converge to a Stackelberg equilibrium.

(2) Experience Replay and Neural Network Training

The middle section of the figure illustrates how the DDPG algorithm employs an experience replay buffer. At each time step, the transition tuple (s_t, a_t, r_t, s_{t+1}) is stored in the buffer. Both the DSO and MG agents sample mini-batches from this buffer to update their neural networks. Each agent maintains two networks: A current network for generating real-time actions and a target network that is updated to stabilize training.

(3) Actor-Critic Structure

On the right side of the figure, the DDPG framework adopts an actor-critic architecture:

The actor network (policy network) takes the current state s_t as input and outputs a deterministic action a_t.

The critic network (value network) evaluates the Q-value $Q_t(s_t, a_t)$, representing the expected cumulative reward.

2025 SIT Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

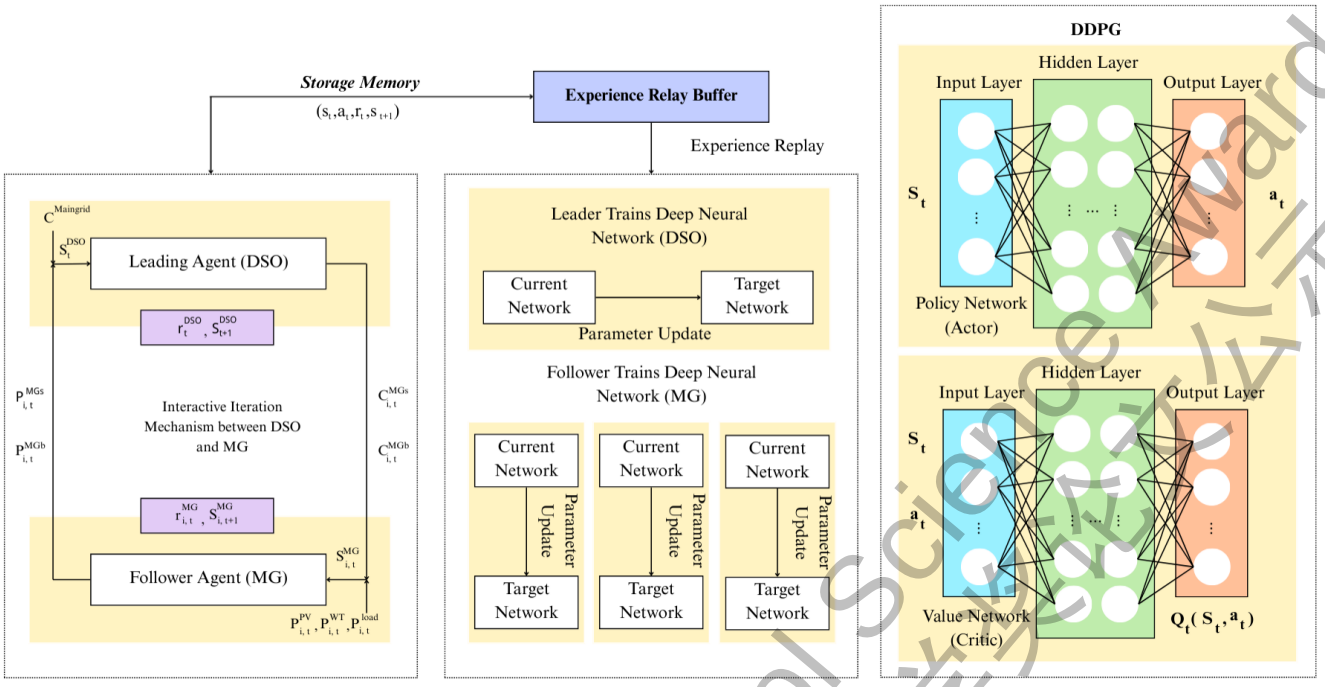

Figure 5: Deep Deterministic Policy Gradient (DDPG) algorithm)

6.6.3 Network Architecture

In our DDPG-based leader-follower framework, each agent implements a four-network architecture: a current policy network (Actor) $\pi(s|\theta_\pi)$, a current value network (Critic) $\pi(s|\theta_Q)$, plus structurally identical target networks $\pi(s|\overline{\theta}_\pi)$ and $\pi(s|\overline{\theta}_Q)$. The target networks use delayed (soft) parameter updates and serve to stabilize temporal-difference targets during training.

The input layer and hidden layer use ReLU activation, while the output layer uses tanh activation.

The proposed framework adopts a dual-role Actor-Critic structure with dual-target networks, carefully designed activation functions, and a deterministic policy to enhance training stability and performance.

(1) Actor Network

Current Policy Network: Interacts with the environment to output optimal actions. Takes the environmental state vector as input and outputs deterministic actions $a_t = \pi(S_t|\theta_\pi)$.

Target Policy Network: Takes the next state S_{t+1} as input and outputs the action for the next time step $a_{t+1} = \pi(S_{t+1}|\overline{\theta}_\pi)$.

(2) Critic Network

Current Critic Network: The current critic network evaluates the action output by the policy network and provides gradient information for policy optimization.

The input is the environment vector S_t and the output action of the current policy network $a_t = \pi(S_t|\theta_\pi)$. The output is the Q value $Q(S_t, a_t|\theta_Q)$.

Target Critic Network: The input is the next state S_{t+1} and the next action output by the target policy network $a_{t+1} = \pi(S_{t+1}|\overline{\theta}_\pi)$.

The output is the target Q value $\overline{Q}(S_{t+1}, \pi(S_{t+1}|\overline{\theta}_\pi)|\overline{\theta}_Q)$.

6.6.4 Training Process

Step 1: Parameter Initialization

2025 S.T. Yau High School Science Award 应用于2025丘成桐中学科学奖

At the beginning of the training process, the parameters of the policy network θ_π and value network θ_Q are randomly initialized.

$$\overline{\theta}_\pi \leftarrow \theta_\pi, \quad \overline{\theta}_Q \leftarrow \theta_Q \tag{68}$$

$$\theta_{\pi,DSO} \leftarrow \{\theta_{\pi,MG}, i \in k\} \tag{69}$$

$$\theta_{Q,DSO}, \left\{\theta_{Q,MG^i} \mid i \in k\right\} \tag{70}$$

Step 2: Action Selection

At each time step t, the DSO observes the current environment state $S_{t,DSO}$ and selects an action $a_{t,DSO}$ using its deterministic policy network. The selected action influences the system and updates the corresponding microgrid (MG) states $S_{i,t,MG}$. Then, the follower agents independently select their own actions $a_{i,t,MG}$ based on their local observations. After executing these actions, both leader and follower agents receive instantaneous rewards $r_{t,DSO}$ and $r_{i,t,MG}$, and the environment transitions to the next states. The following transition tuples are stored into the experience replay buffer:

$$\{S_{t,DSO}, a_{t,DSO}, r_{t,DSO}, S_{t+1,DSO}\} \tag{71}$$

$$\{S_{i,t,MG}, a_{i,t,MG}, r_{i,t,MG}, S_{i,t+1,MG}\} \tag{72}$$

Step 3: Gaussian Noise Injection

To encourage sufficient exploration during training, Gaussian noise is added to the deterministic action output by the actor network.

$$A_t = \pi_t(S_t|\theta_\pi) + N \tag{73}$$

where $\pi_t(S_t|\theta_\pi)$ denotes the output of the policy network at time step t. $\pi_t(S_t|\theta_\pi)$ denotes the output of the policy network at time step t.

The Gaussian noise N follows a normal distribution with mean 0 and variance σ_t^2:

$$N \sim \mathcal{N}(0|\sigma_t^2) \tag{74}$$

where $\sigma_t = \Theta^{-\xi \cdot t}$, Θ represents the initial exploration variance, and ξ is the decay rate.

Step 4: Current Network Update

The minimum loss function is defined as: $L(\theta_Q) = E[q_t - Q(S_t, a_t|\theta_Q)]^2$

where $q_t = r_t + \gamma \max \overline{Q}(S_{t+1}, \pi(S_{t+1}|\overline{\theta}_\pi)|\overline{\theta}_Q)$

The parameter θ_Q is updated using gradient descent to minimize the loss function. The update formula is:

$$\theta_Q \leftarrow \theta_Q - \eta_Q \nabla_{\theta_Q} L(\theta_Q) \tag{75}$$

where η_Q is the learning rate of the value network.

$\nabla_{\theta_Q} L(\theta_Q)$ is the gradient of the loss function, approximated using the function below:

$$\frac{1}{F} \sum_t [\nabla_{a_t} Q(S_t, a_t)|a_t = \pi(S_t|\theta_\pi)] \tag{76}$$

Here, F is the number of samples used for estimation. $\nabla_{a_t} Q(S_t, a_t)$ represents the gradient of the Q value from the value network, indicating the direction that maximizes Q. $\nabla_{\theta_\pi} \pi(S_t)$ denotes the gradients of the policy network.

Step 5: Soft Update of Target Networks

$$\overline{\theta}_\mu \leftarrow \tau\theta_\mu + (1-\tau)\overline{\theta}_\mu \tag{77}$$

$$\overline{\theta}_Q \leftarrow \tau\theta_Q + (1-\tau)\overline{\theta}_Q \tag{78}$$

Where $\tau \ll 1$, meaning the update is carried out in the direction to the greatest extent possible.

2025 S.-T. Yau High School Science Award 仅用于2025丘成桐中学科学奖论文公示

7 Simulation Verification

7.1 Simulation Parameter Settings

The simulation is conducted on the IEEE 33-bus distribution system, which serves as a benchmark for evaluating the performance of the proposed control strategy. The main parameters and configurations are carefully selected to ensure a realistic and comprehensive test environment. The system base voltage is set at 12.66 kV, while the base capacity is 10 MVA. To maintain stable operation, the permissible voltage range is defined between 0.95 p.u. and 1.05 p.u., ensuring that the voltage levels remain within acceptable limits under various operating conditions.

Three microgrids are integrated into the system, with their connection nodes located at buses 12, 23, and 28. These microgrids incorporate a diverse array of components, including wind turbines, photovoltaic (PV) systems, micro gas turbines, energy storage systems, and electrical loads. This configuration allows for a detailed analysis of the interactions between different energy sources and the overall grid stability. It is assumed that the operational constraints and cost coefficients of the three microgrids are identical, simplifying the comparative analysis while still providing valuable insights into the system's behavior.

For the control strategy, two deep neural networks (DNN-A and DNN-C) are employed, each consisting of an input layer, hidden layers, and an output layer. All layers are fully connected, with each hidden layer containing 256 neurons. This architecture ensures sufficient complexity to capture the intricate dynamics of the system. The control framework is based on the Deep Deterministic Policy Gradient (DDPG) algorithm, which is well-suited for continuous action spaces and can effectively handle the challenges posed by the multi-agent environment. The training parameters are meticulously tuned to optimize the learning process and achieve optimal performance in the simulation.

This setup provides a robust foundation for verifying the effectiveness of the proposed control strategy, enabling a thorough evaluation of its capabilities in managing complex distribution systems with multiple microgrids.

Table 2: Parameters of DDPG

Parameter	Value
Actor network learning rate η^{π}	0.0001
Critic network learning rate η^{Q}	0.0005
Actor network noise decay rate ξ	0.001
Critic network discount factor γ	0.9
Target network soft update coefficient τ	0.001
Experience replay buffer capacity B	10000
Experience replay mini-batch sample size F	256
Maximum training episodes E	500

7.2 Training Process Analysis

The evolution of cumulative reward during the offline training phase of the Deep Deterministic Policy Gradient (DDPG) algorithm is presented in Figure 6, alongside comparative results from the standard Deep Q-Network (DQN) and its variant, Dueling-DQN. In the early training phase, the agent incrementally acquires knowledge of the environmental dynamics through interaction, enabling progressive refinement of its decision-making policy via parameter updates in the underlying neural networks. This iterative process facilitates convergence toward an optimal control

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖公示

strategy.

As illustrated in the learning curves, Dueling-DQN exhibits a pronounced improvement in cumulative reward after approximately 150 episodes, followed by stabilization, suggesting effective decoupling of state-value and advantage estimation in its architecture. In contrast, DDPG demonstrates a delayed but substantial performance leap around the 250th episode, ultimately achieving the highest asymptotic reward across all evaluated methods. The standard DQN, while exhibiting steady improvement, converges to a suboptimal policy, as evidenced by its lower final reward. This comparative analysis underscores the superiority of DDPG in continuous control tasks, particularly in capturing nuanced action-value relationships and maintaining long-term performance stability. The abrupt performance gain observed in DDPG likely signifies the successful internalization of critical system dynamics—such as temporal dependencies in energy supply and demand—and the subsequent refinement of control policies through policy gradient updates.

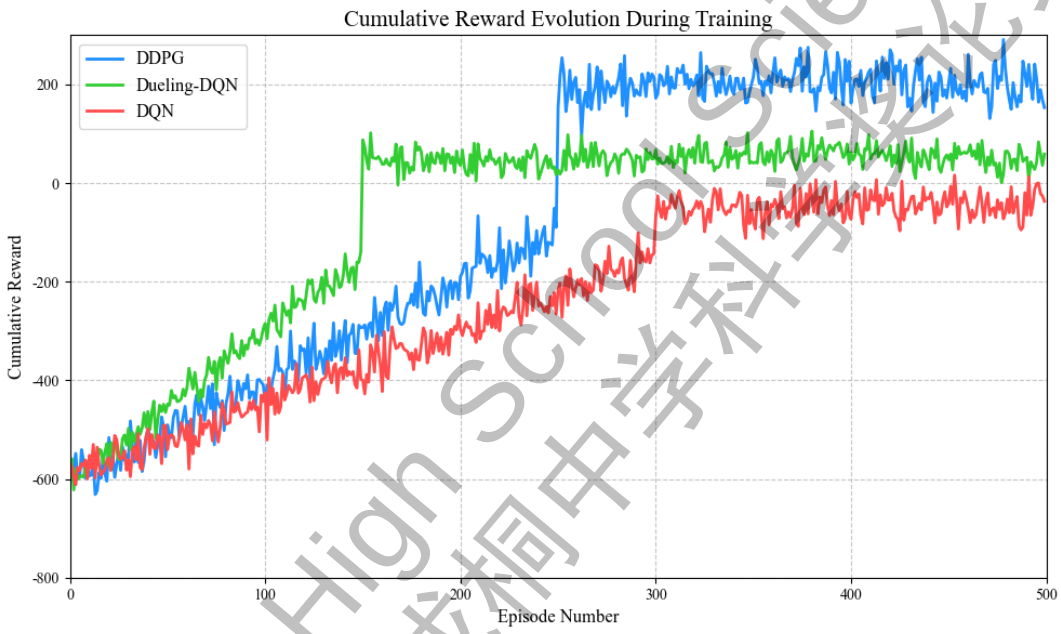

Figure 6: Cumulative Reward Evolution During Training

7.3 Online Operation Performance

To assess the real-time operational efficacy of the trained model, the optimized control strategy was deployed using a representative 24-hour operational dataset encompassing realistic load and renewable generation profiles. The resulting power trajectories for the three interconnected microgrids (MG1, MG2, and MG3) are depicted in Figure.7 and Figure.8, illustrating the model's ability to manage inherent uncertainties in wind and solar generation while meeting time-varying load demands.

The results demonstrate that the proposed Stackelberg-based deep reinforcement learning framework enables effective coordination between the distribution system operator and distributed microgrid entities. Energy storage systems within each microgrid are strategically operated to absorb surplus renewable generation during periods of high production and discharge during supply deficits or peak demand intervals, thereby enhancing local energy autonomy and reducing reliance on external grid support. Furthermore, the temporal alignment between generation, storage, and consumption reflects a coherent optimization of economic and operational objectives.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

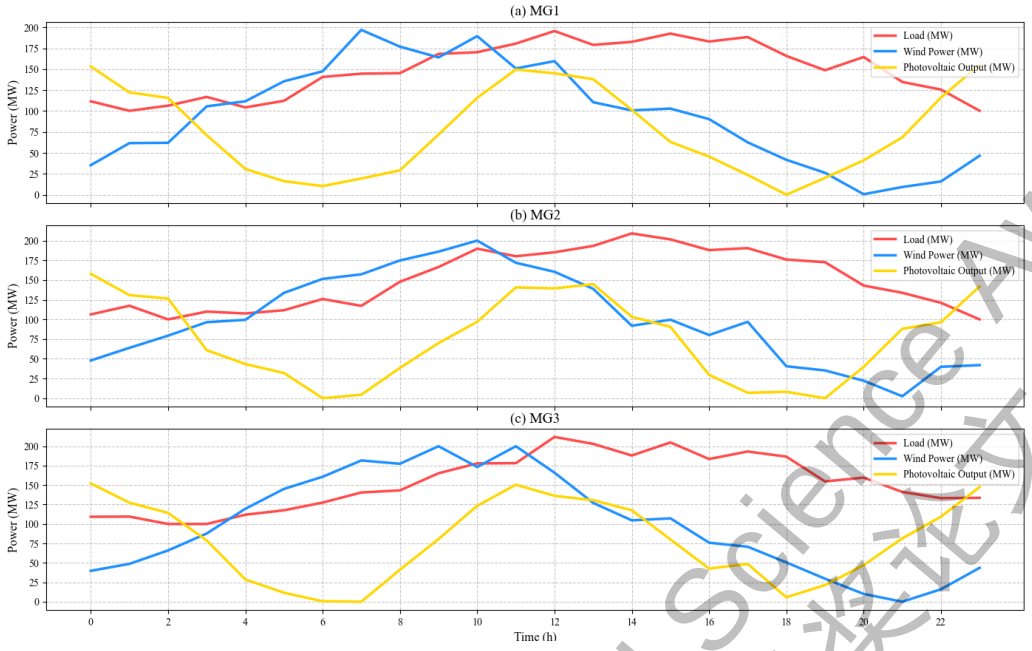

Figure 7: The resulting power trajectories for the three interconnected microgrids

Additionally, the convergence behavior of the agent's performance metric over successive training episodes is shown in Figure.9, where a consistent upward trend followed by stabilization indicates the attainment of a robust and generalizable policy. This convergence suggests that the agent has successfully learned to balance exploration and exploitation, adapting to dynamic grid conditions without requiring explicit system models or re-optimization at each time step. The integration of hierarchical decision-making with model-free reinforcement learning thus enables scalable, adaptive, and computationally efficient energy management in multi-agent distribution networks.

The operation of energy storage systems (ESS) within the three microgrids plays a pivotal role in enhancing system flexibility and ensuring supply-demand balance under fluctuating renewable generation and load conditions. As illustrated in the power profiles, the ESS units are actively engaged in temporal energy shifting—strategically storing excess energy during periods of high renewable output and low demand, and releasing stored energy during generation deficits or peak consumption intervals. This dynamic dispatch mechanism effectively mitigates power imbalances at the point of common coupling, reduces reliance on external grid support, and contributes to voltage and frequency stability within the local networks.

Moreover, the coordinated charging and discharging cycles across MG1, MG2, and MG3 reflect an optimized utilization of distributed storage resources in response to both local and system-wide signals. The control strategy enables predictive scheduling based on forecasted generation and load patterns, while also accommodating real-time deviations through fast-responding reinforcement learning decisions. By leveraging the storage assets as controllable power sources or sinks, the microgrids achieve a higher degree of energy autonomy and resilience against intermittency.

The temporal distribution of ESS operation also reveals the economic rationality embedded in the learned policy. Charging predominantly occurs during off-peak hours or high-generation periods when electricity prices or opportunity costs are lower, whereas discharging is scheduled to coincide with high-price or high-demand intervals, thereby minimizing operational costs and maximizing economic benefits for each microgrid participant. This price-responsive behavior, autonomously learned through the hierarchical reinforcement learning framework, aligns individual microgrid objectives with the overall network-level optimization goals, such as loss minimization

2025 S.T. Yau High School Science Award 仅用于2025丘成桐中学科学奖论文公示

and load flattening.

Furthermore, the depth and duration of charge/discharge events are modulated to preserve battery health and operational constraints, including state-of-charge limits and ramping capabilities. The resulting ESS dispatch profile demonstrates not only effective power balancing but also adherence to technical feasibility, reflecting a well-generalized policy that accounts for both dynamic performance and long-term sustainability. This intelligent coordination underscores the potential of data-driven, multi-agent control architectures in enabling scalable and efficient energy management in modern distribution systems with high penetration of distributed energy resources.

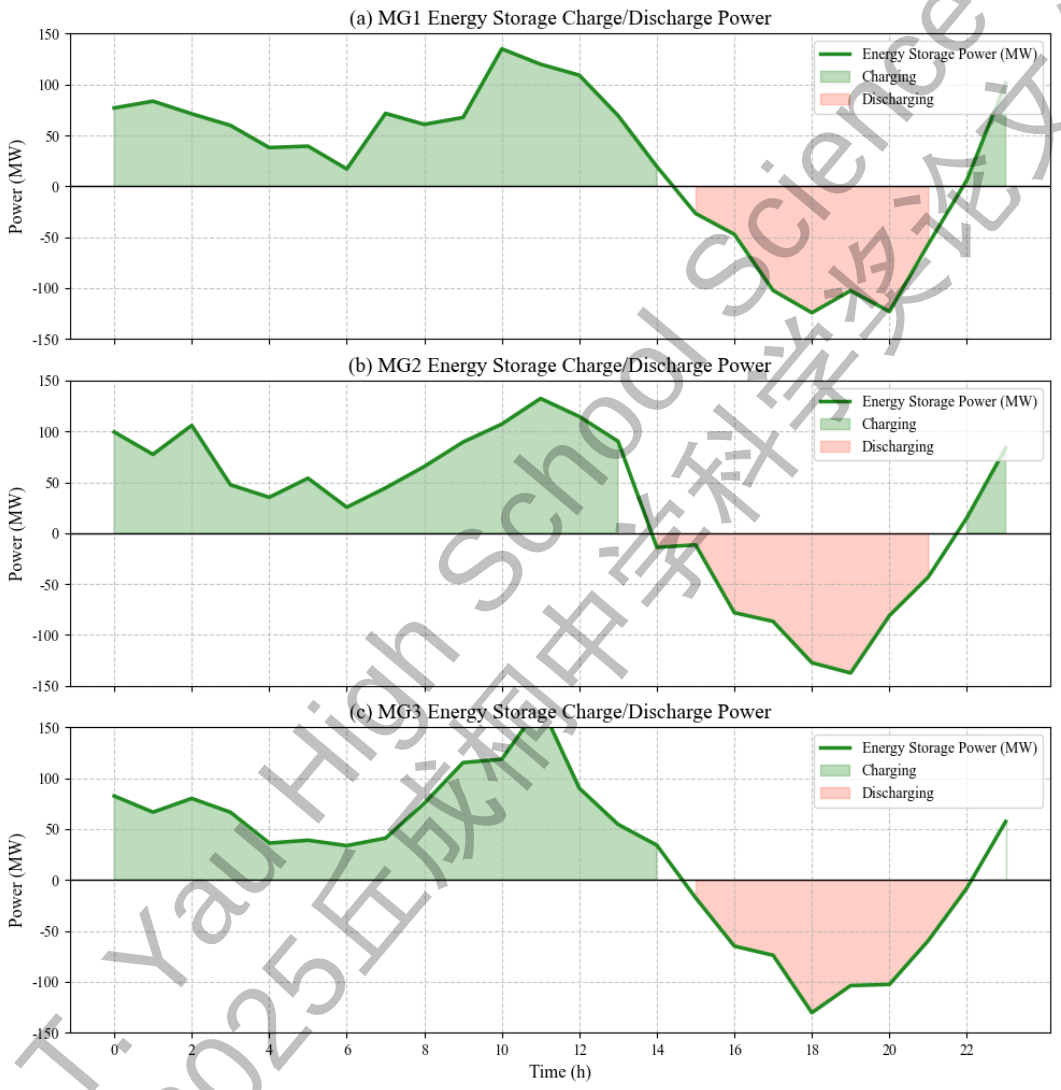

Figure 8: The charging and discharging power of the energy storage systems (ESS) in MG1, MG2, and MG3

The performance trajectory over the training process in Figure.9 exhibits a consistent upward trend followed by asymptotic stabilization, indicating the progressive refinement and eventual convergence of the control policy. This behavior reflects the agent's successful learning of the underlying system dynamics, including the stochasticity of renewable generation, load fluctuations, and operational constraints of energy storage and power exchange. The converged policy demonstrates robust generalization, enabling adaptive decision-making under diverse operating conditions without reliance on explicit re-optimization. The smooth and stable convergence characteristics suggest effective training with minimal oscillation or performance degradation, attributable to well-designed algorithmic components such as experience replay, target networks, and reward shaping.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

The resulting policy supports real-time, model-free execution with low computational latency, while the hierarchical framework facilitates coordinated decision-making among multiple entities, reflecting the emergence of an effective and scalable control strategy for distributed energy systems.

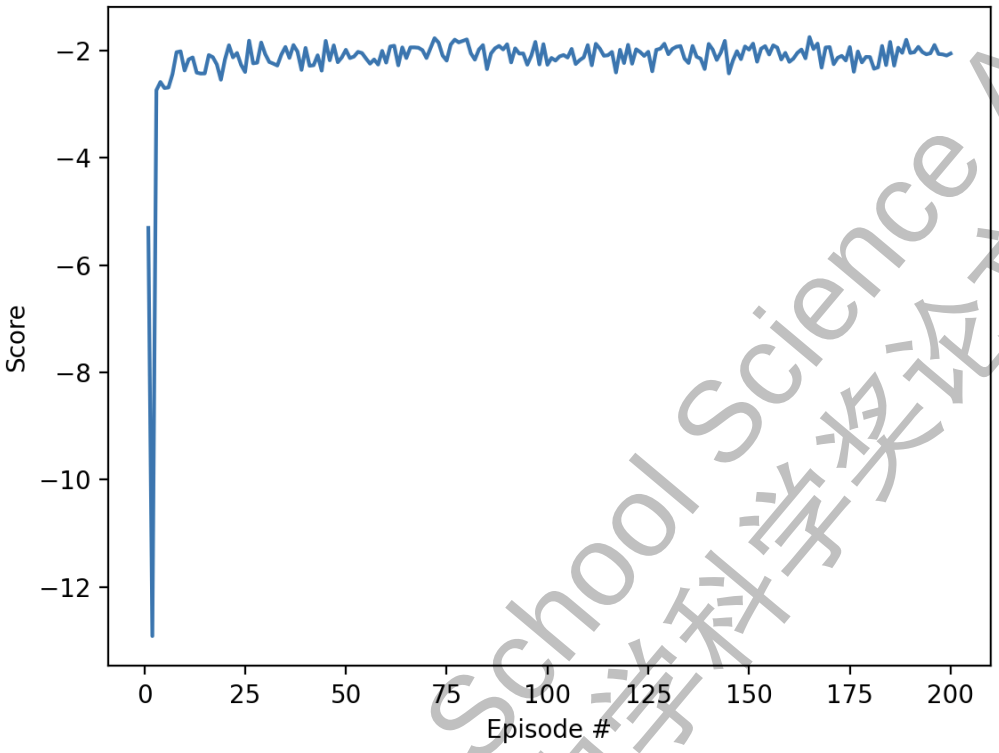

Figure 9: Learning progress

The cumulative reward trajectories of the leader (distribution network operator) and the follower agents (microgrids MG1, MG2, and MG3) illustrate the strategic interaction inherent in the hierarchical control framework. In the early training stages, significant fluctuations are observed in the reward profiles of all agents, indicative of an exploratory phase where policies are repeatedly adjusted in response to evolving system conditions and inter-agent interactions. These variations arise from the coupled decision-making structure, in which the leader issues coordination signals—such as price incentives or power exchange targets—while the followers optimize their local operations subject to these constraints. As training proceeds, the amplitude of oscillations diminishes, and the reward sequences converge toward steady values, suggesting that the system reaches a stable operating regime. This convergence reflects a consistent alignment between the leader's coordination strategy and the followers' response behaviors, resulting in a coherent control structure that simultaneously satisfies individual and system-wide objectives, such as cost efficiency, demand satisfaction, and network stability.

Concurrently, the training loss dynamics of the actor and critic networks provide insight into the numerical stability and learning efficacy of the underlying reinforcement learning mechanism. The critic loss decreases rapidly and remains at a low level, indicating accurate and consistent estimation of the value function, which is crucial for guiding policy updates. In contrast, the actor loss initially fluctuates due to policy exploration but gradually increases and stabilizes, reflecting a systematic shift toward control policies that yield higher long-term returns. This behavior is consistent with the objective of maximizing cumulative reward through iterative policy improvement. The eventual stabilization of both losses, with minimal divergence, demonstrates a well-balanced interaction between policy evaluation and policy optimization. The convergence characteristics confirm the

2025 S.T.Yau High School Science Award 仅用于2025丘成桐中学科学奖论文公示

robustness of the learning process in a high-dimensional, continuous control setting, supported by architectural and algorithmic choices such as target networks, experience replay, and appropriate regularization. The overall training behavior validates the framework's capacity to derive effective, decentralized control strategies for multi-agent energy systems through data-driven optimization.

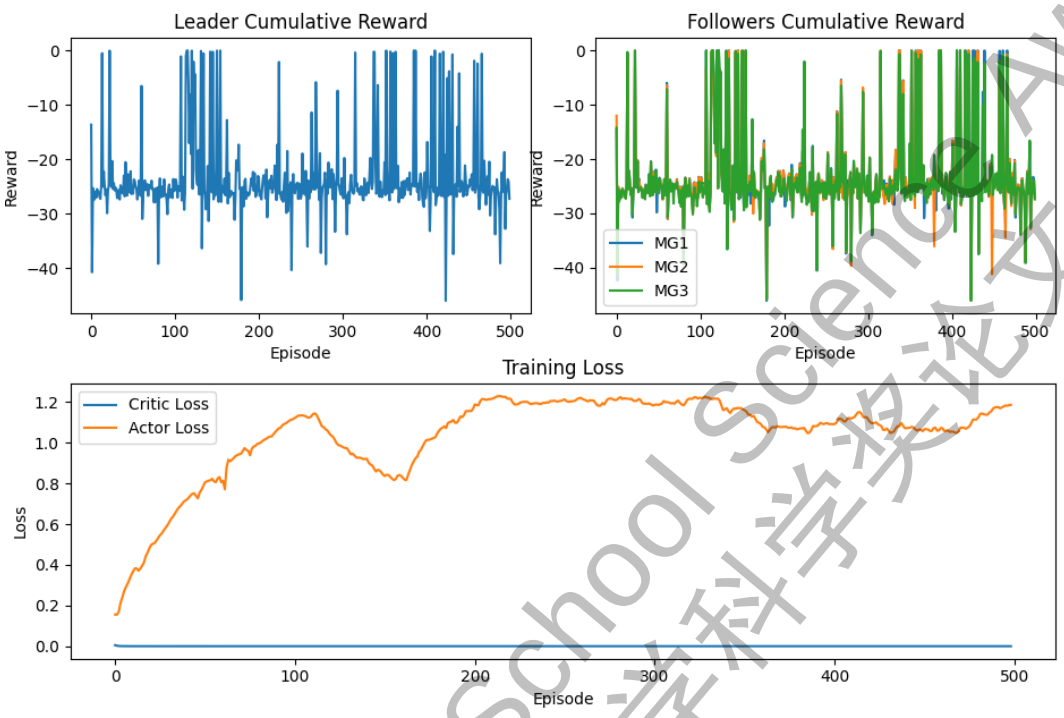

Figure 10: The cumulative rewards of the leader and the followers

8 Online Operation Analysis

The proposed prioritized dueling double deep Q-network (Per Dueling-DDQN) framework demonstrates significant advantages over conventional methods in the context of multi-microgrid energy management, particularly in terms of operational economy, computational efficiency, privacy preservation, and robustness under uncertainty. In economic performance, the method achieves a cost-effective scheduling strategy by leveraging the hierarchical structure of the Stackelberg game, which enables the distribution system operator (DSO) to guide microgrid behaviors through incentive-based signals such as energy pricing or power exchange targets. Compared to standard deep Q-network (DQN) approaches, the dueling architecture enhances value estimation accuracy by decoupling state value and action advantage functions, thereby facilitating more precise policy evaluation and improved decision quality. This refinement leads to superior long-term economic outcomes, as the agent learns to balance immediate operational costs with future system states under stochastic renewable generation and load variations. When benchmarked against traditional mathematical programming techniques—such as mixed-integer second-order cone programming (MISOCP) and model predictive control (MPC)—the proposed method demonstrates enhanced economic efficiency, particularly in dynamic environments where re-optimization frequency and model inaccuracies limit the performance of model-based approaches.

From a computational standpoint, the framework enables real-time decision-making with minimal online computational burden, a critical requirement for fast-varying power systems. Traditional optimization methods typically require solving complex, non-convex problems at each

2025 S.T. Yau High School Science Award
仅用于2025年成同中学科学奖论文公示

decision interval, resulting in substantial computational latency that hinders scalability and responsiveness. In contrast, the deep reinforcement learning approach shifts the computational load to an offline training phase, where the policy is learned through interaction with a simulated environment. Once trained, the policy can be deployed online with negligible inference time, allowing for immediate responses to changing grid conditions without iterative solving or reliance on precise system models. This model-free characteristic further enhances practicality, as it eliminates the need for accurate parameterization of system dynamics, which are often difficult to obtain in real-world applications.

A key innovation of the proposed framework lies in its distributed architecture, which inherently supports privacy-preserving coordination. Conventional centralized methods, such as MISOCP reformulations of equilibrium problems, necessitate full access to private operational data—including generation costs, load profiles, and storage constraints—from all participating microgrids, raising significant concerns regarding data confidentiality and operational autonomy. In contrast, the Stackelberg-based DRL framework operates on a signal-based interaction mechanism: only high-level coordination signals, such as energy prices and power exchange setpoints, are exchanged between the DSO and microgrids, while sensitive local information remains isolated within individual agents. This design aligns with practical regulatory and commercial constraints in decentralized energy markets, where entities are unwilling or unable to share proprietary data. The integration of experience replay with prioritized sampling further enhances learning efficiency and stability, enabling the agent to focus on high-impact transitions and accelerate convergence. The resulting policy demonstrates strong generalization capabilities, maintaining robust performance under diverse and uncertain operating conditions, including sudden load changes, renewable intermittency, and equipment variability. These attributes collectively establish the framework as a scalable, adaptive, and operationally viable solution for real-time energy management in multi-microgrid systems.

9 Discussion

This study presents a hierarchical, distributed optimization framework that effectively addresses the challenges of coordination, privacy, and uncertainty in multi-microgrid systems by integrating Stackelberg game theory with deep reinforcement learning. The method successfully models the strategic interaction between the distribution system operator and microgrids as a leader–follower game, enabling decentralized decision-making while ensuring alignment between local and system-wide objectives. By eliminating the need for full information exchange, the approach preserves data privacy and reduces communication overhead, making it suitable for practical deployment in heterogeneous and independently operated energy networks. Furthermore, the model-free nature of the algorithm allows it to adapt to complex, non-linear system dynamics without relying on explicit mathematical models, thereby enhancing robustness in the face of modeling inaccuracies and environmental stochasticity.

Despite these advancements, several avenues exist for further strengthening the framework. First, as system scale increases, the dimensionality of the state space grows significantly, potentially leading to slower convergence and instability during training. Incorporating advanced feature extraction techniques, such as autoencoders or principal component analysis, or employing dimensionality reduction methods could improve learning efficiency and enable scalability to larger networks with numerous distributed energy resources. Second, the current formulation assumes simplified behavioral models for flexible loads and distributed generation units. More sophisticated modeling of demand elasticity, storage degradation, and diverse resource dynamics would better capture the true operational flexibility of microgrids and lead to more realistic and effective control

2025 S-T. Yau High School Science Award
仅用于2025年丘成桐中学科学奖公示

strategies. Third, the transferability of the learned policy across different network topologies or operational regimes remains limited. Integrating transfer learning or meta-learning techniques could enable the agent to rapidly adapt to new environments or configurations with minimal retraining, thereby enhancing generalization and reducing deployment costs. Additionally, incorporating risk-sensitive objectives or robust optimization layers could further improve resilience under extreme uncertainty or adversarial conditions.

10 Conclusion

This study proposes a novel hierarchical energy management framework for multi-microgrid systems by combining Stackelberg game theory with a deep reinforcement learning architecture based on Per Dueling-DDQN. The approach models the distribution system operator as a leader and multiple microgrids as followers, capturing the strategic interplay inherent in decentralized power systems. Through this structure, the method achieves coordinated optimization of energy scheduling while respecting the autonomy and privacy of individual entities. Extensive simulations demonstrate that the proposed framework outperforms conventional methods—including standard DQN, MISOCP, and MPC—in terms of operational economy, computational efficiency, and adaptability to uncertain environments. By enabling real-time, model-free decision-making with minimal online computation, the approach is well-suited for dynamic grid operations characterized by high renewable penetration and fluctuating demand. The distributed design ensures that sensitive operational data remain localized, with only high-level coordination signals exchanged between layers, thus addressing critical privacy and scalability challenges in modern energy systems. Future enhancements through advanced state representation, refined resource modeling, and transfer learning are expected to further improve the method's practicality and robustness. Overall, this work highlights the potential of integrating game-theoretic principles with deep reinforcement learning to develop intelligent, scalable, and resilient control strategies for the next generation of distributed power systems.

References

[1] Krishnamurthi R V, Moran A E, Forouzanfar M H, et al. The global burden of hemorrhagic stroke: a summary of findings from the GBD 2010 study[J]. Global heart, 2014, 9(1): 101-106.

[2] Tiili P, Lehto M, Halminen O, et al. Hemorrhagic stroke in atrial fibrillation: Trends in incidence, case fatality, and prior oral anticoagulation[J]. Journal of the American Heart Association, 2025, 14(12): e040360. DOI:10.1161/JAHA.120.040360. https://doi.org/10.1161/JAHA.124.040360

[3] Parisio A, Rikos E, Glielmo L. A model predictive control approach to microgrid operation optimization[J]. IEEE Transactions on Control Systems Technology, 2014, 22(5): 1813-1827.

[4] Shayan M E, Najafi G, Ghobadian B, et al. Multi-microgrid optimization and energy management under boost voltage converter with Markov prediction chain and dynamic decision algorithm[J]. Renewable Energy, 2022, 201: 179-189.

[5] Shuai H, Li F, Pulgar-Painemal H, et al. Branching dueling Q-network-based online scheduling of a microgrid with distributed energy storage systems[J]. IEEE Transactions on Smart Grid, 2021, 12(6): 5479-5482.

2025 S.T. Yau High School Science Award 仅用于2025丘成桐中学科学奖公示

[6] Li H, Yang Y, Liu Y, et al. Federated dueling DQN based microgrid energy management strategy in edge-cloud computing environment[J]. Sustainable Energy, Grids and Networks, 2024, 38: 101329.

[7] Dong X, Li X, Cheng S. Energy management optimization of microgrid cluster based on multi-agent-system and hierarchical Stackelberg game theory[J]. IEEe Access, 2020, 8: 206183-206197.

[8] Liu N, Yu X, Wang C, et al. Energy sharing management for microgrids with PV prosumers: A Stackelberg game approach[J]. IEEE Transactions on Industrial Informatics, 2017, 13(3): 1088-1098.

[9] Li B, Zhao R, Lu J, et al. Energy management method for microgrids based on improved Stackelberg game real-time pricing model[J]. Energy Reports, 2023, 9: 1247-1257.

[10] Hu C, Cai Z, Zhang Y, et al. A soft actor-critic deep reinforcement learning method for multi-timescale coordinated operation of microgrids[J]. Protection and Control of Modern Power Systems, 2022, 7(1): 29.

[11] Thrun S, Schwartz A. Issues in using function approximation for reinforcement learning[C]//Proceedings of the 1993 connectionist models summer school. Psychology Press, 2014: 255-263.

[12] Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning[C]//International conference on machine learning. PMLR, 2016: 1995-2003.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

11 Acknowledgements

This research was completed under the patient guidance of my supervisor, Ms. Wen, combined with my own persistent efforts. I would like to express my deepest gratitude to Ms. Wen for her selfless support throughout the entire research process. I also sincerely thank my school for providing students with space and resources for independent exploration.

The inspiration for this project came from my interest in China's national "carbon peak and carbon neutrality" policy. During my studies, I learned that with the increasing integration of renewable energy sources such as solar and wind power, traditional power dispatching methods are facing new challenges. How multiple microgrids can collaborate efficiently and trade electricity fairly has become a practical yet complex issue. At first, I only had a vague idea of studying energy optimization. It was Ms. Wen who suggested that I look into the intersection of smart grids and artificial intelligence. By reviewing university-level open courses and academic papers, I gradually realized the potential of deep reinforcement learning in addressing dynamic decision-making problems. This led me to the idea of applying this technology to microgrid scheduling, and with Ms. Wen's encouragement, I formally established my research direction.

The entire project was independently carried out by me, including topic selection, literature review, model design, programming implementation, simulation testing, and paper writing. Ms. Wen is a science and technology mentor at my school and provided guidance for this competition completely on a voluntary, unpaid basis. Although she did not directly participate in coding or data processing, she played a crucial role in helping me understand fundamental concepts such as "microgrids" and "reinforcement learning." During the data preparation phase, I used publicly available datasets from the U.S. National Renewable Energy Laboratory (NREL) for wind and solar power generation, along with IEEE standard distribution system configurations. Due to inconsistencies in the original data formats, I encountered difficulties in processing them. Ms. Wen taught me how to use Python for data cleaning and normalization, which helped me understand the importance of data preprocessing in real-world research.

The modeling phase was the most challenging part. I attempted to build a multi-agent system using the MADDPG algorithm, but the program failed to converge, and the reward values fluctuated wildly. For a period of time, I spent nearly every day tuning parameters and even considered giving up. It was Ms. Wen who reminded me to pay attention to mechanisms such as experience replay and target network updates, and suggested that I refer to stabilization techniques used in related research papers. After multiple attempts, I implemented prioritized experience replay and soft target updates, which eventually stabilized the training process. She also advised me to incorporate multiple objectives—such as economic cost, carbon emissions, and supply-demand balance—into the reward function, making the model more practically meaningful.

When writing the paper, I lacked experience in academic writing. My first draft was disorganized and poorly expressed. Ms. Wen carefully reviewed it paragraph by paragraph, pointed out logical flaws, helped me restructure the sections, and taught me how to describe technical details accurately. She particularly emphasized the need to clearly explain why I chose a certain method and how it differs from traditional approaches, which deepened my understanding of the research.

This project has taught me that scientific research is not something that can be achieved overnight, but rather a process of continuous trial, error, and improvement. I am deeply grateful to Ms. Wen for her consistent encouragement and guidance throughout this journey.

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖公开

LUCIA YANG/杨静姝

INTELLIGENT / ENERGETIC / OPEN-MINDED

INFORMATION

Birthday: 2010.04.06
Gender: Female
Nationality: Chinese
Living place: Shanghai

CONTACT

School:Vanke School Pudong
Grade: G10
Position: Class President
E-mail:luuuucia@yeah.net

EDUCATION

Primary School
2016.9 – 2021.6(G1-G5)

Shanghai Vanke Bilingual School-VKBS
https://vkbs.dtd-edu.cn/en/

Secondary School
2021.9 – (G6- G8)

Shanghai Starriver Bilingual School-SSBS
https://www.ssbs.sh.cn/

High School
2024.9 – (G9-)

Vanke School Pudong
https://vsp.dtd-edu.cn/

AWARDS

- Academic Excellence Award in school year from 2021 to 2025
- 1st Prize Award, China Region, **IMMC** 2025
- Team Leader&2nd Prize Award@Asia Region, 2nd Prize Award@China, **ISSDC** 2025(International Space Settlement Design Competitions) 2025
- Distinguished Honor Roll Award(1%) in 2023 **AMC 10**
- Scored 9/15 points in American Invitational Mathematics Examination(**AIME**) 2024
- Global Gold award in **Physical Bowl** D1 2024
- Achieved Gold level of **USACO**
- Global Gold award in UKMT Itermediate Mathematical Challenge 2023 & 2024
- Scored 100 points in TOFEL 2023
- Second Prize Award in Shanghai Youth Orchestral Music Contest(Harp)
- Gold Metal of Original Oratory Contest in National Speech and Debate Association China(NSDA) 2023
- 2nd Prize Award in Shanghai Division of China Youth Music Competition

ACTIVITIES

- Class President in school year from 2022 to 2025
- Founder&President of the **Logics&Reasons Club**, to promote the beauty of logics and mathmetical reasoning in school
- Recomposed Alexander Hamilton Musicals and performed during School's New Year Show 2023
- Led a team of 6 classmates to win the 1st Prize Award of Physics and BioChemicals in school's Science Fair 2024.
- Head of Publicity Department of the Student Union in 2023
- Co-founded the Girls' Basketball Club of SSBC with G8 English Teacher
- Harp palyer in the orchestra at school

HOBBIES

- Basketbal
- Traveling
- Playing Chess&Puzzles
- Harp
- Drawing
- Reading

SELF EVALUATION

- A free spirit with a open mind and kind heart
- Excellent academic performance in school
- Gifted in mathematics
- Easy-going and congenial, with a strong sense of responsibility and good team-spirit.
- Active in school and social activities with strong communication skills and organizational capability.

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

Haiyang Wen (Serena)

17812767230

Haiyangwen16@gmail.com

Education

2011-2015 University of Pittsburgh, Pittsburgh, PA USA

- **Bachelor of Science in Mathematics**
- **Bachelor of Science in Economics**

Working Experience

2022.8-current
Vanke Shanghai School, Pudong

Mathematics Teacher & Homeroom Teacher & AP Assistant Coordinator

- CIE IGCSE Mathematics Teaching for Grade 9 and Grade 10.
- IB DP Mathematics AA Teaching for Grade 11 and Grade 12.
- AP Precalculus, Calculus AB/BC, Computer Science, Statistics Teacher.
- AP Coordinator duties.
- Math Competition guidance: AMC10, AMC12, ARML and Euclid.
- Homeroom teacher duties.

2021.8-2022.8
ULink Shanghai International School

Mathematics & Physics Teacher

- CIE IGCSE, Pure Mathematics, Mechanics and Further Mathematics.
- Math Competition guidance: AMC10, AMC12 and Euclid.
- MAT and STEP preparation.

2019.11- 2021.8
Vision Overseas Consulting (Shanghai) Cp.,Ltd

AP Calculus AB & Calculus BC; Pure Math and Further Math; IB Application & Interpretation (11-12 grade); IB Analysis & Approaches (11-12 grade); A Level AS & A2 Math; Computing for academic purpose

- Developing the first-year college preparation course: Introduction to Python and Introduction to Java.
- Tutoring majority computer science examinations, such as AP/IGCSE Computer Science, covering Primitive Types, Using Objects, Boolean Expressions and if Statement, Iteration, Writing Classes, etc.

2016.03-2019.11
Wholeren Education LLC (Pittsburgh, PA USA)

Senior Math & Physics & Computer Science Instructor

- 3-years tutored experience in sophomore, junior and senior math courses and computing courses around the U.S Universities.
- Developing computing lesson plans for high school curriculum and college.

Teaching Subject

- Mathematics
- Physics

Skillset

- Fluent English in writing and speaking

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示