Journal of
*Marine Science
and Engineering*

MDPI

Article

Dual−Layer Distributed Optimal Operation Method for Island Microgrid Based on Adaptive Consensus Control and Two−Stage MATD3 Algorithm

Zhibo Zhang [1,2,*], Bowen Zhou [1,2,*], Guangdi Li [1,2], Peng Gu [1,2], Jing Huang [3] and Boyu Liu [4]

[1] College of Information Science and Engineering, Northeastern University, Shenyang 110819, China; liguangdi@mail.neu.edu.cn (G.L.); gupeng@mail.neu.edu.cn (P.G.)
[2] Key Laboratory of Integrated Energy Optimization and Secure Operation of Liaoning Province, Northeastern University, Shenyang 110819, China
[3] State Grid Electric Power Research Institute Wuhan Efficiency Evaluation Company Limited, Wuhan 430072, China; huangjing1@sgepri.sgcc.com.cn
[4] School of Electrical Engineering and Telecommunications, UNSW Sydney, Sydney, NSW 2052, Australia
[*] Correspondence: 2100690@stu.neu.edu.cn (Z.Z.); zhoubowen@ise.neu.edu.cn (B.Z.)

Abstract: Island microgrids play a crucial role in developing and utilizing offshore renewable energy sources. However, high operation costs and limited operational flexibility are significant challenges. To address these problems, this paper proposes a novel dual−layer distributed optimal operation methodology for islanded microgrids. The lower layer is a distributed control layer that manages multiple controllable distributed fuel−based microturbines (MTs) within the island microgrids. A novel adaptive consensus control method is proposed in this layer to ensure uniform operating status for each MT. Moreover, the proposed method can achieve the total output power of MTs to follow the reference signal provided by the upper layer while ensuring plug−and−play capability for MTs. The upper layer is an optimal scheduling layer that manages various forms of controllable distributed power sources and provides control reference signals for the lower layer. Additionally, a two−stage twin−delayed deterministic policy gradient (MATD3) algorithm is utilized in this layer to minimize the operating costs of island microgrids while ensuring their safe operation. Simulation results demonstrate that the proposed methodology can effectively reduce the operating costs of island microgrids, unify the operational status of MTs, and achieve plug−and−play capability for MTs.

Keywords: island microgrid; offshore renewable energy; optimized scheduling; consensus control; deep reinforcement learning

check for updates

Citation: Zhang, Z.; Zhou, B.; Li, G.; Gu, P.; Huang, J.; Liu, B. Dual−Layer Distributed Optimal Operation Method for Island Microgrid Based on Adaptive Consensus Control and Two−Stage MATD3 Algorithm. *J. Mar. Sci. Eng.* **2023**, *11*, 1201. https://doi.org/10.3390/jmse11061201

Academic Editor: Yassine Amirat

Received: 15 May 2023
Revised: 3 June 2023
Accepted: 7 June 2023
Published: 9 June 2023

1. Introduction

Microgrid technology is an effective solution for addressing the problem of insufficient local power supply while enhancing the utilization of renewable energy resources. Compared to onshore microgrids, island microgrids possess a more abundant renewable energy supply [1]. However, harsh maritime climate conditions expose the equipment in island microgrids to humid, saline, and moldy operating environments. Such environments can result in higher operating costs and a higher probability of damage to power equipment used in island microgrids [2]. In addition, the high percentage of renewable energy supply also makes the island power system more vulnerable and reduces the security of island microgrid operations [3]. However, given the importance of islands in marine resources and safety, it is crucial to ensure the safety and flexibility of island microgrid operation. Therefore, this study proposes a novel dual−layer distributed optimization operation method for island microgrids based on adaptive consensus control and a two−stage MATD3 algorithm to enhance the operational flexibility and economic efficiency of island microgrids. We hope this study can provide a valuable reference for the future planning and development of island microgrids.

Although microgrids incorporate various energy sources, such as solar and wind, microturbines (MTs) remain a primary energy source for ensuring normal microgrid operation [4]. Island microgrids often deploy multiple distributed MTs to achieve safety and flexibility. Owing to its simplicity and speed, the centralized control structure is most commonly used for multiple distributed MTs in microgrids. Reference [5] proposed a centralized control architecture for microgrids by utilizing a phase−locked loop to obtain the system frequency and compared it with the rated frequency to determine the total power deficit. Accordingly, the power reference instruction for each distributed power source was obtained based on this power deficit. In [6], an adaptive droop strategy was proposed for microgrids based on a centralized control scheme that could compensate for the impact caused by the voltage drop of the feeder to improve the reactive power distribution accuracy. However, the centralized control scheme depended on the central controller and communication lines. Any failure to these components might disrupt the control of distributed MTs, resulting in abnormal microgrid operation. The distributed control method based on consensus control provides an alternative approach to control distributed MTs without relying on a central controller. In such methods, each distributed MT communicates with adjacent distributed MTs to achieve control of all distributed MTs. Moreover, it does not affect the operation of the microgrid system, even during partial communication loss [7]. Therefore, the distributed control method based on consensus control is more suitable for use in harsh environmental conditions on island microgrids to improve their safety and flexibility. In ref. [8], a cloud−edge collaboration−based distributed control method was proposed which could alleviate the tremendous computational pressure caused by excessively centralized computation tasks while solving the optimal control strategy for the power grid. However, this method required the construction of a cloud−based service platform, which was expensive to build. Refs. [9,10] achieved the plug−and−play function of distributed MT with guaranteed frequency recovery and optimal tide of the island microgrid, respectively. However, neither of them considered the operational economics of the island microgrid. In ref. [11], a dual−layer consensus control method was proposed for a multi−microgrid that achieves capacity−based allocation of MT output power within and between microgrids. Based on this architecture, ref. [12] adopted a new consensus control method with an equal micro−increment rate in the upper layer control to reduce the operational cost of distributed power sources. However, the operating costs of other devices in the microgrid were not considered; therefore, further research is needed to determine whether the method could reduce the overall operating cost of the microgrid.

Generally, the optimal scheduling problem for island microgrids is a mixed−integer nonlinear programming problem (MINLP) [13]. Classical optimization methods [14], planning−based methods [15,16], and heuristic algorithms [17–19] are commonly employed to solve such a problem. However, the environmental conditions of island microgrid operation tend to be complex, and the amount of data that algorithms need to handle is large, making it difficult for the above method to achieve efficient optimal scheduling of microgrids in real time [20]. With the development of machine learning algorithms, deep reinforcement learning algorithms have demonstrated good adaptability in solving grid problems with complex models [21]. For instance, ref. [22] improved the double deep Q network (DDQN) algorithm by using convolutional neural networks (CNNs) and multiple buffer zones, which greatly improved the learning ability of the algorithm. However, the action space of the algorithm was discrete, and the accuracy of the optimal strategy obtained often depended on the degree of discretization of the action space. In ref. [23], a model−actor−critic reinforcement learning neural network architecture combined with event−triggered control was constructed. This architecture could significantly accelerate the learning speed of the reinforcement learning neural network while reducing the computational and communication requirements of the control process. In ref. [24], the deep deterministic policy gradient (DDPG) algorithm was used to provide effective scheduling strategies for household microgrids. However, it overlooked the overestimation problem that might occur when the DDPG algorithm updated iteratively. In ref. [25], the MT and energy storage were simultaneously controlled by using an improved soft actor−critic (SAC) algorithm. However, since MT and energy storage were two different energy forms, using the same neural networks and deep reinforcement learning parameters would adversely affect the learning efficiency of the SAC algorithm. A summary of the existing methods for optimal operation of island microgrids is shown in Table 1.

Table 1. Summary of the existing methods for optimal operation of island microgrids.

Ref. No	Method	Advantages	Disadvantages
[5]	Centralized control	Guaranteed power distribution by capacity	Failure of the central controller will cause the whole system to be abnormal
[6]	Adaptive droop strategy	Compensates for the impact caused by the voltage drop of the feeder to improve the reactive power distribution accuracy	Failure of the central controller will cause the whole system to be abnormal
[8]	Cloud−edge collaboration	Alleviates the tremendous computational pressure caused by excessively centralized computation tasks	Requires the construction of a cloud−based service platform, which is expensive to build
[9]	Multi−agent system based multi−layer architecture	Achieves the plug−and−play function of distributed MT with guaranteed frequency recovery	No consideration of the operational economics of island microgrids
[10]	Distributed control using two sensors	Achieves the plug−and−play function of distributed MT with guaranteed optimal tide	No consideration of the operational economics of island microgrids
[11]	Dual−layer consensus control	Achieves capacity−based allocation of MT output power within and between microgrids	No consideration of the operational economics of island microgrids
[12]	Equal micro−increment dual−layer consensus control	Reduces the operational cost of distributed power sources	No consideration of the operating costs of other devices in the microgrid
[22]	Improved DDQN algorithm	Greatly improves the learning ability of the algorithm	The action space of the algorithm is discrete, and the computational accuracy is low
[23]	Model−actor−critic reinforcement learning combined with event−triggered control	Reduces the computational and communication requirements of the control process	The calculation process is more complicated
[24]	DDPG algorithm	Provides effective scheduling strategies for household microgrids	Overlooks the overestimation problem that might occur when the DDPG algorithm updated iteratively
[25]	Improved SAC algorithm	Provides effective scheduling strategies for microgrids	The training process can adversely affect the learning efficiency of the SAC algorithm

To the best of our knowledge, although there have been some studies on the optimal operation of island microgrids, they mostly consider island microgrids containing a single MT. In fact, few studies have been conducted on the optimal operation methods for island microgrids comprising multiple distributed MTs. Therefore, this paper proposes a novel dual−layer distributed optimization operation method for island microgrids containing multiple distributed MTs. The proposed method combines consensus control and multi−agent reinforcement learning algorithms and reasonably improves them. Consequently, the proposed method can improve the effectiveness of optimal scheduling decisions while improving the economy and flexibility of island microgrid operation. The main contributions of this paper include:

1. A two−layer distributed optimal operation framework is established for island microgrids, which develops upon the operating environment model of an island microgrid. The lower layer is a distributed control layer, and it uses the consensus control method with the goal of unifying the operating status of distributed MTs. The upper layer is the optimal scheduling layer, and it aims to maximize the economic benefits of island microgrids by using a two−stage MATD3 algorithm.
2. A novel adaptive consensus control method is proposed in the lower layer, which allocates the output power of distributed MTs according to their capacities while ensuring that the total output power of the MTs follows a reference signal provided by the upper layer. Additionally, the proposed method guarantees the plug−and−play capability of distributed MTs, which means that the above functionalities are maintained even when distributed MTs are plugged in or plugged out from the system.
3. A two−stage MATD3 is proposed in the upper layer, which can maximize the operational economy of island microgrids by adjusting the reference signals of distributed MTs and energy storage. Moreover, the method incorporates a pre−training stage to enhance the training effectiveness of the algorithm, while also mitigating the sensitivity of the MATD3 algorithm to the parameters, thereby reducing the complexity of parameter tuning.

The remainder of the paper is organized as follows. Section 2 presents the modeling of the islanded microgrid. Section 3 introduces a dual−layer distributed optimization operation method for the islanded microgrid. Section 4 provides a simulation analysis of the proposed method. Finally, conclusions are drawn in Section 5.

2. Island Microgrid Model

Generally, an island microgrid includes both land−based energy and sea−based energy. In this context, the island microgrid designed in this study consists of photovoltaic (PV) power generation, wind turbine (WT) power generation, MTs, tidal power generation, wave power generation, energy storage (ES), and loads, as depicted in Figure 1.

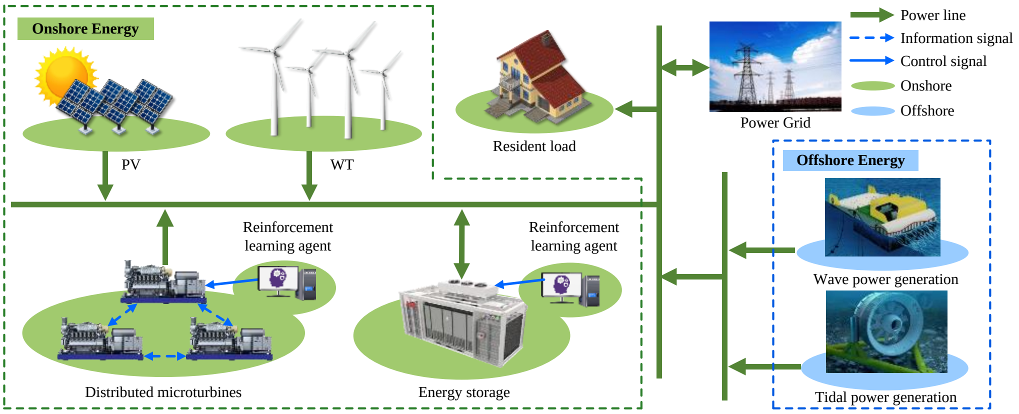


Figure 1. Structure of the island microgrid designed in this study.

First, we model the distributed MTs, ES, and busbar of the island microgrid.

- Distributed MTs

MTs provide an adjustable power supply to island microgrids by combusting fuels, which can effectively reduce the dependence of island microgrids on external grids. Assuming there are m distributed MTs in the island microgrid, the total operational cost can be expressed as in the form of a quadratic function as [26]

$$C_{\mathrm{MT}}(t) = \sum_{i=1}^{m} [\alpha_i (P_i(t))^2 + \beta_i P_i(t) + c_i], \tag{1}$$

where $C_{\text{MT}}(t)$ denotes the total fuel cost of the distributed MTs during the t period; $P_i(t)$ represents the power output of the $i-$th MT during the t period; a_i, β_i, and c_i are the fuel cost coefficients of the $i-$th MT.

The output power of MTs should meet the following constraint:

$$P_{i,\min} < P_i(t) < P_{i,\max}, \tag{2}$$

$$-R_{i,\text{down}} \leq P_i(t) - P_i(t-1) \leq R_{i,\text{up}}, \tag{3}$$

where $P_{i,\min}$ and $P_{i,\max}$ denote the maximum and minimum output power of the $i-$th MT, respectively, while $R_{i,\text{down}}$ and $R_{i,\text{up}}$ denote the upward ramp and downward ramp constraints of the $i-$th MT, respectively.

- Energy storage

Energy storage can coordinate with renewable energy sources with randomness and fluctuations, playing a role in "peak shaving and valley filling" to ensure the reliability and economy of microgrids. Considering the charging and discharging power and the state of charge (SOC) of the energy storage, the charging and discharging process of energy storage can be expressed as [20]

$$SOC(t+1) = \begin{cases} SOC(t) + \frac{\eta P_{\text{ES}}(t)\Delta t}{S_{es}}, & P_{\text{ES}}(t) > 0, \\ SOC(t) + \frac{P_{\text{ES}}(t)\Delta t}{\zeta S_{es}}, & P_{\text{ES}}(t) < 0, \\ SOC(t), & P_{\text{ES}}(t) = 0, \end{cases} \tag{4}$$

where $SOC_i(t)$ represents the state of charge of the energy storage at time t; $P_{\text{ES}}(t)$ represents the power output or absorption of ES during the t period; η and ζ denote the charging and discharging efficiencies of the ES device, respectively; S_{es} indicates the rated capacity of the ES.

The cost of ES is composed of two main components, namely capacity cost and power cost, which can be expressed as [27]

$$C_{\text{ES}}(t) = g_E S_{es} + g_P |P_{\text{ES}}(t)|, \tag{5}$$

where $C_{\text{ES}}(t)$ represents the total of energy storage during the t period; g_E indicates the capacity cost coefficient of ES; g_P denotes the power cost coefficient of ES.

The charging/discharging power and SOC of ES should meet the following constraints:

$$\begin{cases} 0 < P_{\text{ES}}(t) < P_{ch.\max}, & P_{\text{ES}}(t) > 0, \\ 0 < |P_{\text{ES}}(t)| < P_{dis.\max}, & P_{\text{ES}}(t) < 0, \end{cases} \tag{6}$$

$$SOC_{\min} < SOC(t) < SOC_{\max} \tag{7}$$

where $P_{ch,\max}$ and $P_{dis,\max}$ represent the maximum power of ES during charging and discharging, respectively; $SOC_{\max}$ and $SOC_{\min}$ denote the maximum and minimum SOC of ES.

- Island microgrid busbar

The busbar of an island microgrid acts as a bridge for energy exchange between the island microgrid and the external grid. Assuming that all renewable energy output power in the island microgrid is integrated into the grid, the busbar of the island microgrid must maintain power balance. Therefore, the power of the busbar of the island microgrid can be expressed as

$$P_{\text{Grid}}(t) = P_L(t) - [\sum_{i=1}^{m} P_i(t) + P_{\text{PV}}(t) + P_{\text{WT}}(t) + P_{\text{ES}}(t) + P_{\text{Tidal}}(t) + P_{\text{Wave}}(t)], \tag{8}$$

where $P_{\text{Grid}}(t)$ represents the power exchanged between the island microgrid and the external grid. $P_{\text{Grid}}(t) > 0$ represents the period when the island microgrid purchases power from the external grid. Conversely, $P_{\text{Grid}}(t) < 0$ represents the period in which the island microgrid sells power to the external grid. $P_{\text{PV}}(t)$ represents the output power of PV generation; $P_{\text{WT}}(t)$ indicates the output power of WTs; $P_{\text{Tidal}}(t)$ indicates the output power of tidal energy generation; $P_{\text{Wave}}(t)$ represents the output power of wave energy generation; $P_L(t)$ represents the power consumed by the load in the island microgrid.

The cost of purchasing electricity from the external grid or the benefit of selling electricity to the island microgrid can be expressed as

$$\begin{cases} C_{\text{Grid}}(t) = \sigma^b(t)P_{\text{Grid}}(t), & P_{\text{Grid}}(t) > 0, \\ C_{\text{Grid}}(t) = \sigma^s(t)P_{\text{Grid}}(t), & P_{\text{Grid}}(t) \le 0, \end{cases} \tag{9}$$

where $C_{\text{Grid}}(t)$ represents the cost of energy exchange between the island microgrid and the external power grid; $\sigma^b(t)$ and $\sigma^s(t)$ represent the electricity prices at which the island microgrid purchases and sells electricity to the external power grid, respectively.

In summary, the total operating cost of the island microgrid proposed in this paper can be expressed as

$$F(t) = \sigma^L P_{\text{L}}(t) + C_{\text{MT}}(t) + C_{\text{ES}}(t) + C_{\text{grid}}(t), \tag{10}$$

where $F(t)$ represents the total operational cost of the island microgrid, while σ^L denotes the electricity price at which the island microgrid sells power to its internal users.

3. Dual−Layer Distributed Optimal Operation Method for Island Microgrids

The proposed two−layer distributed optimal operation framework for island microgrids is illustrated in Figure 2. The upper layer contains two different deep reinforcement learning agents, namely the MT and ES agents, for computing the optimal scheduling policy for the island microgrid. The MT agent offers reference signals for the distributed MTs, while the ES agent offers reference signals for ES. The lower layer contains multiple distributed MTs and ES devices. The multiple distributed MTs achieve output power distribution by capacity and total output power following the reference signal via mutual communication and the consensus control method. Notably, as long as one or a few MTs in the lower layer can receive the reference signal from the upper layer, control over all MTs can be achieved.

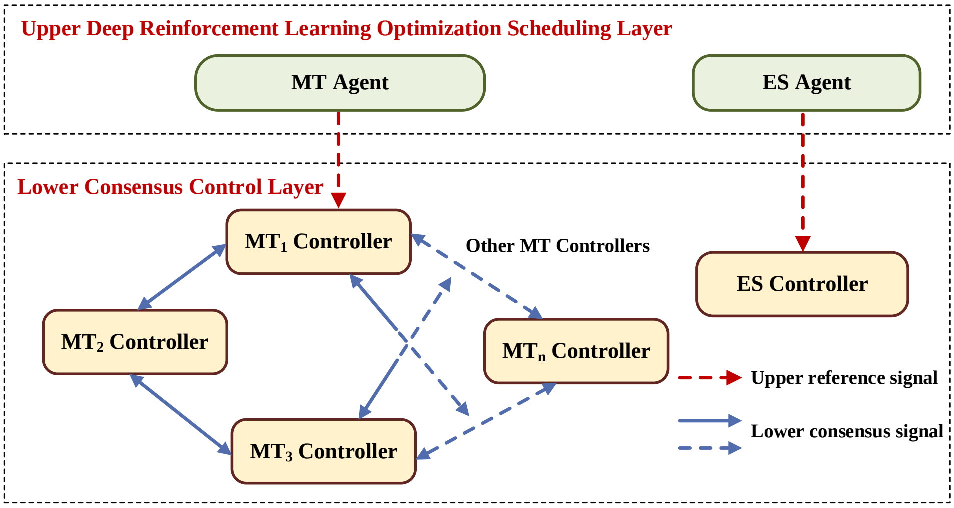


Figure 2. Two−layer distributed optimal operation framework for the island microgrid.

3.1. Lower Layer Control Method

3.1.1. Fundamental Theory of Multi−Agent Consensus Control

Each controller of the MT in the island microgrid can be regarded as a consensus agent, and the communication relationships among multiple consensus agents can be represented by graph $G_\varepsilon(V_\varepsilon,\ \psi_\varepsilon,\ K_\varepsilon,\ B_\varepsilon)$. Assuming there are n_ε agents in the graph, $V_\varepsilon = \{V_{\varepsilon,\,1}\cdots,\ V_{\varepsilon,\,n_\varepsilon}\}$ represents the set of nodes, each of which represents a consensus agent. $\psi_\varepsilon \in V_\varepsilon \times V_\varepsilon$ represents the set of edges, representing the communication lines between nodes. $K_\varepsilon = (K^\varepsilon_{ij})_{(n_\varepsilon-1)\times(n_\varepsilon-1)}$ represents the weights of edges. If there is a communication connection between $V_{\varepsilon,i} \in V_\varepsilon$ and $V_{\varepsilon,j} \in V_\varepsilon$, then $k^\varepsilon_{ij} > 0$, otherwise, $k^\varepsilon_{ij} = 0$. $B_\varepsilon = \mathrm{diag}(k^\varepsilon_{1,0},\cdots,\ k^\varepsilon_{n_\varepsilon,0})$ represents the leading adjacency matrix. If $V_{\varepsilon,i} \in V_\varepsilon$ can receive a reference signal, then $k^\varepsilon_{i0} > 0$, otherwise, $k^\varepsilon_{i0} = 0$. Assuming that each node has a scalar state signal x_i, each node can update its state based on its own state and the state signal of the nodes it communicates with. Based on the consensus control scheme, the rules for updating the state of the node have the following two forms [28]:

$$\dot{x}_i(t) = \sum_{j\in V_\varepsilon} k^\varepsilon_{ij}(x_j(t) - x_i(t)), \tag{11}$$

$$\dot{x}_i(t) = \sum_{j\in V_\varepsilon} [k^\varepsilon_{ij}(x_j(t) - x_i(t)) + k^\varepsilon_{i0}(x_{ref} - x_i(t))] \tag{12}$$

where $\dot{x}_i$ denotes the differential of the state variable x_i. According to [28], if the communication network graph among consensus agents has a spanning tree, then the following two theorems hold.

Theorem 1. *If the update rule defined by (11) is employed, then the states of all agents will converge to a consensus value. Specifically, if the communication network graph is balanced, then the states of all agents will converge to the average of their initial states, i.e.,*

$$\lim_{t\to\infty} x_i(t) = [\sum_{i\in V_\varepsilon} x_i(0)]/n_\varepsilon. \tag{13}$$

Theorem 2. *If the update rule defined by (12) is employed, then the states of all agents will converge to the reference value x_{ref}, i.e.,*

$$\lim_{t\to\infty} x_i(t) = x_{ref}. \tag{14}$$

The proof process of the two theorems mentioned above can be found in [28]. Notably, the reference value x_{ref} can also possess dynamics.

3.1.2. Lower Layer Adaptive Consensus Control for Island Microgrids

The objective of the lower layer adaptive consensus control scheme for the island microgrid is to allocate the output power of multiple distributed MTs in the island microgrid based on their capacities. Moreover, the lower layer control scheme ensures that the total output power of the remaining running distributed MTs is equal to the reference value provided by the upper layer. Assuming there are m distributed MTs in the island microgrid, and the communication network graph between MTs is balanced, the control objectives of the lower layer can be described as

$$\lim_{t\to\infty} |P_i(t)/P_{i,\max} - P_j(t)/P_{j,\max}| = 0, \tag{15}$$

$$\lim_{t\to\infty} |\sum_{i=1}^{m} P_i(t) - P_{ref}(t)| = 0 \tag{16}$$

where P_i and P_j denote the output power of the $i-$th and $j-$th MTs, respectively; $P_{i,\max}$ and $P_{j,\max}$ denote the maximum output power of the $i-$th and $j-$th MTs; P_{ref} denotes the power reference signal provided by the upper layer.

Three state signals, the number state signal $n_i(t)$, capacity ratio state signal $\eta_i(t)$, and the power state signal $P_i(t)$, are transmitted among the distributed MTs via mutual communication. Among them, the initial value of the number state signal of one of the MTs is set to 1, while the initial value of the number state signal of other MTs is set to 0. The updating formula for the number state of each MT is

$$\dot{n}_i = \sum_{j=1}^{m} k_{ij}^{n}(n_j - n_i), \tag{17}$$

where $\dot{n}_i$ denotes the differential of n_i. According to Theorem 1, (17) can converge the number state values of each MT to the average value of the initial number state values of all MTs, which is the reciprocal of the number of MTs in the island microgrid, i.e.,

$$\lim_{t\to\infty} n_i(t) = [\sum_{i=1}^{m} n_i(0)]/m = 1/m. \tag{18}$$

The initial value of the capacity ratio state $\eta_i(0) = P_{i,\max}/P_{\text{nom}}$, where P_{nom} indicates the standard value of MT capacity of the island microgrid. The updating formula for the capacity ratio state of each MT is

$$\dot{\eta}_i = \sum_{j=1}^{m} k_{ij}^{\eta}(\eta_j - \eta_i), \tag{19}$$

where $\dot{\eta}_i$ denotes the differential of η_i. According to Theorem 1, (19) can converge the state value of the capacity ratio of each MT to the initial average value of the capacity ratios of all MTs, i.e.,

$$\lim_{t\to\infty} \eta_i(t) = [\sum_{i=1}^{m} \eta_i(0)]/m. \tag{20}$$

From (15), it is evident that when the power of the island microgrid reaches equilibrium, the output power of each MT needs to satisfy

$$\frac{P_1}{P_{1,\max}} = \frac{P_2}{P_{2,\max}} = \cdots = \frac{P_m}{P_{m,\max}}. \tag{21}$$

By combining (21) with the expression for the capacity ratio, it can be deduced that:

$$\frac{P_1}{\eta_1(0)} = \frac{P_2}{\eta_2(0)} = \cdots = \frac{P_m}{\eta_m(0)}. \tag{22}$$

Let $P_{ref}^{m} = P_i/\eta_i(0)$, then $P_i = \eta_i(0) \cdot P_{ref}^{m}$. As expressed in (15), when the power of the island microgrid reaches equilibrium, the sum of the output power of all MTs needs to equal the reference signal provided by the upper layer, that is

$$P_{ref} = P_1 + P_2 + \cdots P_m = \eta_1(0)P_{ref}^{m} + \eta_2(0)P_{ref}^{m} + \cdots + \eta_m(0)P_{ref}^{m} = [\sum_{i=1}^{m} \eta_i(0)] \cdot P_{ref}^{m}, \tag{23}$$

By combining (18), (20), and (23), we obtain the following expression:

$$P_{ref}^{m} = P_{ref} \cdot \frac{n_i(t)}{\eta_i(t)}. \tag{24}$$

Considering P_{ref}^{m} as the output power reference signal of the MT capable of obtaining upper−layer reference signals, and letting $P_{k,i} = P_i/\eta_i$, the updating formula for the power state of each MT is

$$\dot{P}_{k,i} = \sum_{j=1}^{m} [k_{ij}^{P}(P_{k,j} - P_{k,i}) + k_{i0}^{P}(P_{ref}^{m} - P_{k,i})], \tag{25}$$

where $\dot{P}_{k,i}$ denotes the differential of $P_{k,i}$. According to Theorem 2, (25) can enable all P_i/η_i of the MTs to converge to P_{ref}^{m}, which achieves power allocation of the MTs' output according to their capacity. It can also be inferred from (23) that the sum of the output power of all MTs is equal to the reference signal P_{ref} provided by the upper layer.

3.1.3. Plug−and−Play Improvements for Adaptive Consensus Control

Unlike the traditional distributed MT consensus control method, the control objective of this study is to make the total output power of all MTs track the reference signal provided by the upper layer. When a new distributed MT is plugged in, the adaptive consensus control scheme described above still enables the total output power of the MTs to follow the reference signal provided by the upper layer. However, when an MT is plugged out, the output power of the remaining MTs will remain unchanged, and then the total output power of the MTs will not be able to follow the reference signal. This is because when an MT is plugged in or plugged out, the communication topology between distributed MTs will be changed. At this time, for the state signals n_i and η_i, the controller will take the state signal value at the moment the MT is plugged in or plugged out as the new initial state value. Therefore, the state value of the MT converges to the average value of new initial state value. Once the MT is plugged in, the average value of the new initial value is still equal to the average of the original initial state values of all MTs in the island microgrid, while it is not equal after the MT is plugged out. As a result, the total output power of the remaining MTs after the MT is plugged out will not be equal to the upper reference signal.

Consider the state signal n_i of the distributed MTs as an example and assume a new MT is plugged in at time t_r. The control objective after the MT is plugged in will be

$$\lim_{t\to\infty} n_i(t) = [\sum_{i=1}^{m+1} n_i(0)]/(m+1). \tag{26}$$

The state value of the MT plugged in at time t_r is its initial state value, i.e., $n_{m+1}(t_r) = n_{m+1}(0)$. Assuming that the system is already balanced before the new MT is plugged in, it can be inferred from Theorem 1 that:

$$n_1(t_r) = n_2(t_r) = \cdots n_m(t_r) = [\sum_{i=1}^{m} n_i(0)]/m. \tag{27}$$

It can be inferred from (27) and Theorem 1 that, after the new MT is plugged in, the system will converge to a new equilibrium state as

$$\begin{aligned} \lim_{t\to\infty} n_i(t) &= \tfrac{n_1(t_r)+n_2(t_r)+\cdots+n_m(t_r)+n_{m+1}(t_r)}{m+1} = \tfrac{\{\sum\limits_{i=1}^{m} n_i(0)]/m\}\cdot m+n_{m+1}(0)}{m+1} \\ &= [\sum_{i=1}^{m+1} n_i(0)]/(m+1). \end{aligned} \tag{28}$$

It can be seen in (28) that the control objectives shown in (29) can be achieved.

However, assuming that at time t_s the $m-$th agent is plugged out, our control objective will be

$$\lim_{t\to\infty} n_i(t) = [\sum_{i=1}^{m-1} n_i(0)]/(m-1). \tag{29}$$

Similarly, assuming that the system is already balanced before the $m-$th MT is plugged out, it can be inferred from Theorem 1 that:

$$n_1(t_s) = n_2(t_s) = \cdots n_m(t_s) = [\sum_{i=1}^{m} n_i(0)]/m \tag{30}$$

According to (30) and Theorem 1, it can be concluded that after the $m-$th MT is plugged out, the system will converge to a new equilibrium state as

$$\begin{aligned} \lim_{t\to\infty} n_i(t) &= \tfrac{n_1(t_s)+n_2(t_s)+\cdots+n_{m-1}(t_s)}{m-1} = \tfrac{\{\sum\limits_{i=1}^{m} n_i(0)]/m\}\cdot(m-1)}{m-1} \\ &= [\sum_{i=1}^{m} n_i(0)]/m \neq [\sum_{i=1}^{m-1} n_i(0)]/(m-1) \end{aligned} \tag{31}$$

According to (31), it can be inferred that when the $m-$th MT is plugged out, the state n_i of the remaining MTs will remain unchanged. Therefore, the control objectives shown in (29) will not be achieved.

Based on the analysis above, we improve on the adaptive consensus control method proposed in Section 3.1.3 that enables the control objective shown in (29) to be achieved after the MT is plugged out. Assuming that when a distributed MT is plugged out, the communication link with other distributed MTs remains conductive for a brief period, enabling the disconnected distributed MT to send its signal to the connected distributed MTs, thereby enabling the remaining MTs to reach a new equilibrium.

We use t_s^- to represent the moment just before t_s, and use t_s^+ to represent the moment just after t_s. First, the value of the system's number of distributed MTs m at the moment just before t_s is obtained as $m = 1/n_m(t_s^-)$. Subsequently, the status value of the unplugged MT changes to the following expression:

$$n_m = \int \sum_{j=1}^{m} k_{ij}^{n}(n_j - n_m)\mathrm{d}t + \frac{m}{m-1}[n_m(t_s) - n_m(0)]. \tag{32}$$

From (32) we can learn that:

$$\begin{cases} n_i(t_s^+) = n_i(t_s^-), & i = 1,\ 2,\ \cdots,\ m-1, \\ n_i(t_s^+) = n_i(t_s^-) + \tfrac{m}{m-1}[n_i(t_s^-) - n_i(0)], & i = m. \end{cases} \tag{33}$$

According to Theorem 1, the new equilibrium state to which the state signal n_i of each MT will converge is

$$\begin{aligned} \lim_{t\to\infty} n_i(t) &= \tfrac{n_1(t_s^+)+n_2(t_s^+)+\cdots+n_{m-1}(t_s^+)+n_m(t_s^+)}{m} \\ &= \tfrac{n_1(t_s^-)+n_2(t_s^-)+\cdots+n_{m-1}(t_s^-)+n_m(t_s^-)+\frac{m}{m-1}[n_m(t_s^-)-n_m(0)]}{m}. \end{aligned} \tag{34}$$

Combined with (33) and (34), the following is obtained:

$$\begin{aligned} \lim_{t\to\infty} n_i(t) &= \tfrac{\{[\sum\limits_{i=1}^{m} n_i(0)]/m\}\cdot m+\frac{m}{m-1}[[\sum\limits_{i=1}^{m} n_i(0)]/m-n_m(0)]}{m} \\ &= \tfrac{\frac{m-1}{m}\sum\limits_{i=1}^{m} n_i(0)+[\sum\limits_{i=1}^{m} n_i(0)]/m-n_m(0)}{m-1} = [\sum_{i=1}^{m-1} n_i(0)]/(m-1). \end{aligned} \tag{35}$$

According to (35), the improved method proposed in this study can achieve the control objective expressed in (29) after the MT is plugged out. Once the state signal reaches a new equilibrium, the MT that was plugged out will cut off all communication links. According to (31), the state signal n_i of the remaining MTs will remain unchanged even after the communication links are cut off. Therefore, the control objective expressed in (29) can still be achieved.

The same control method as n_i is used for state signal η_i. In summary, the lower−layer controller of the island microgrid can be classified as

$$\begin{cases} N = \frac{1}{n_i(t_{i,s}^-)}, \\ P_{ref}^m = P_{ref} \cdot \frac{n_i(t)}{\eta_i(t)}, \\ n_i = \int \sum\limits_{j=1}^m k_{ij}^n (n_j - n_i)\mathrm{d}t + l_i\{\frac{N}{N-1}[n_i(t_{i,s}^-) - n(0)]\}, \\ \eta_i = \int \sum\limits_{j=1}^m k_{ij}^\eta (\eta_j - \eta_i)\mathrm{d}t + l_i\{\frac{N}{N-1}[\eta_i(t_{i,s}^-) - \eta(0)]\}, \\ P_{k,i} = \int \sum\limits_{j=1}^m [k_{ij}^P (P_{k,j} - P_{k,i}) + k_{i0}^P (P_{ref}^m - P_{k,i})]\mathrm{d}t, \end{cases} \tag{36}$$

where $t_{i,s}^-$ denotes the moment just before the $i-$th distributed MT is plugged out; N indicates the number of distributed MTs at $t_{i,s}^-$. l_i denotes the sign function indicating whether the $i-$th distributed MT is plugged out; that is, if the $i-$th distributed MT is plugged out, $l_i = 1$, otherwise, $l_i = 0$.

3.1.4. Stability Analysis of Adaptive Consensus Control in the Lower Layer of the Island Microgrid

For convenience of notation, let $n = (n_1, n_2, \cdots, n_m)^T$, $\eta = (\eta_1, \eta_2, \cdots, \eta_m)^T$, $P = (P_1, P_2, \cdots, P_m)^T$, and $P_k = (P_{k,1}, P_{k,2}, \cdots, P_{k,m})^T$, and the adjacency matrices for the three state signals transmitted in the MT communication network graph are given as $A_n = (k_{ij}^n)_{m\times m}$, $A_\eta = (k_{ij}^\eta)_{m\times m}$, and $A_P = (k_{ij}^P)_{m\times m}$, respectively. The leading adjacency matrix for the state variable P is $B = (k_{i0}^P)_{m\times m}$, and the Laplacian matrices for the three communication quantities are L_n, L_η, and L_P. For convenience, let $\varepsilon \in \{n, \eta, P\}$. Accordingly, $L_\varepsilon = D_\varepsilon - A_\varepsilon$, where $D_\varepsilon = \mathrm{diag}(d_1^\varepsilon, d_2^\varepsilon, \cdots, d_m^\varepsilon)$, and $d_i^\varepsilon = \sum\limits_{j=1}^m k_{ij}^\varepsilon$.

According to (36), the lower−layer controller can be formulated as follows:

$$\begin{cases} \dot{n} = -L_n n, \\ \dot{\eta} = -L_\eta \eta, \\ \dot{P}_k = -(L_P + B)P_k + BP_{ref}^m, \\ P_{ref}^m = P_{ref} n \cdot / \eta\ . \end{cases} \tag{37}$$

Subsequently, the Lyapunov candidate can be expressed as

$$V(n, \eta, P_k) = n^T n + \eta^T \eta + P_k^T P_k. \tag{38}$$

Based on the previous statement that $n_1, n_2 \cdots n_m \le 1$, differentiating along the trajectory of the system (37), we have

$$\dot{V}\Big|_{(37)} \le -[2\lambda_2(L_n) - \kappa]\, n^T n - [2\lambda_2(L_\eta) - \kappa]\eta^T\eta - [2\lambda_{\min}(L_P + B) - \frac{2P_{ref}^2}{\kappa \cdot \eta^T \eta}\lambda_{\max}(B^2)]P_k^T P_k, \tag{39}$$

where $\lambda_{\min}$ denotes the smallest eigenvalue, λ_2 denotes the second smallest eigenvalue, and κ denotes an arbitrary positive constant. Therefore, a sufficient condition for $\dot{V}\Big|_{(37)} < 0$ is

$$\frac{P_{ref}^2 \lambda_{\max}(B^2)}{\lambda_{\min}(L_P + B)\eta^T \eta} < \kappa < 4\min\{\lambda_2(L_n),\ \lambda_2(L_\eta)\}. \tag{40}$$

From (40), we have

$$P_{ref}^2 \lambda_{\max}(B^2) < 4\lambda_{\min}(L_P + B)\min\{\lambda_2(L_n),\ \lambda_2(L_\eta)\}\eta^T \eta. \tag{41}$$

Therefore, we can conclude that the stability of the lower control scheme of the island microgrid can be ensured when L_n, L_P, B, and P_{ref} of the communication network between the MTs satisfy (41).

3.2. Upper Layer Optimal Scheduling Method for the Island Microgrid Based on Two−Stage MATD3 Algorithm

3.2.1. Markov Model for Optimal Scheduling of the Island Microgrid

The upper−layer optimal scheduling of the island microgrid adopts a multi−agent deep reinforcement learning method. The decision−making process of deep reinforcement learning can be described as a Markov decision process (MDP) [29], which is generally composed of five elements, namely $\{S,\ a,\ P_{S,S'},\ r,\ \gamma\}$. Specifically, S represents the state space, which is a set of environmental state information observable by the agent; a represents the action space, which is the set of actions taken by the agent; $P_{s,s'}$ represents the state transition probability, which indicates the probability that the environment will transition from state S to state S' after the agent takes action a; r represents the immediate reward, which indicates the reward given to the agent by the environment after taking action a in state S; and γ represents the discount factor, which reflects the impact of the action taken at the current time on the agent's reward in the future. The state space, action space, and reward function of the island microgrid discussed in this paper are designed as follows.

- State space

The state space refers to a collection of environmental information observable by a deep reinforcement learning agent. The state space of the island microgrid designed in this paper includes the operating time, user load, wind power generation, photovoltaic power generation, tidal power generation, wave power generation, microgrid bus power, MT power, ES power, ES SOC, and time−of−use electricity price. The MTs and ES are controlled by the MT agent and ES agent, respectively, and the two agents observe different environmental state variables. Considering the fuel supply problem of the MTs, we arrange the distributed MTs in the inland areas of the island. The state variables observed by the MT agent include the total output power of the MTs, PV power generation, wind power generation, user load, and external grid time−of−use of the electricity price. Considering that energy storage devices can mitigate power fluctuations in the grid and ensure the safe integration of the island microgrid, we set up ES devices in the coastal areas of the island. The state variables observed by the ES agent include the system operating time, ES power, ES SOC, microgrid busbar power, tidal power generation, and wave power generation. Therefore, the state space of the MT agent and the ES agent can be described as

$$S_{\mathrm{MT}} = [P_{\mathrm{MT,sum}},\ P_{\mathrm{PV}},\ P_{\mathrm{WT}},\ P_{\mathrm{L}},\ \sigma^b,\ \sigma^s], \tag{42}$$

$$S_{\mathrm{ES}} = [t,\ P_{\mathrm{ES}},\ SOC,\ P_{\mathrm{grid}},\ P_{\mathrm{Tidal}},\ P_{\mathrm{Wave}}], \tag{43}$$

where S_{MT} denotes the state space of the MT agent; S_{ES} indicates the state space of the ES agent; $P_{\mathrm{MT,sum}}$ indicates the total output power of the MTs, which can be obtained from (23) and (24) that

$$P_{\mathrm{MT,sum}} = (P_i \cdot \eta_i)/[\eta_i(0) \cdot n_i]. \tag{44}$$

From (44), it can be inferred that the MT agent only needs to be connected to the controller of one of the MTs to observe the total output power of all MTs.

- Action space

In the microgrid designed in this study, the actions that can be controlled include the output power of MTs and the charging/discharging power of ES. Therefore, the action spaces of the MT and ES agents can be defined as

$$A^{\mathrm{MT}} = [P_{ref}^{m}], \tag{45}$$

$$A^{\mathrm{ES}} = [P_{\mathrm{ES},\ ref}], \tag{46}$$

where A^{MT} indicates the action space of the MT agent; A^{ES} denotes the action space of the ES agent; and $P_{\mathrm{ES},\ ref}$ denotes the reference signal for the charging/discharging power of the energy storage.

- Reward function

When a deep reinforcement learning agent selects an action, the environment will provide a reward. However, if the action chosen by the agent causes the operational state of the island microgrid to exceed the environmental constraints, the environment will impose a penalty. In this study, the environmental constraint penalties originate from the ramp constraint of MTs and the SOC constraint of ES. The penalty expressions are

$$C_{i,\mathrm{MT}}^{c} = -\lambda_{i,\mathrm{MT}}^{c} \cdot \max\{P_i(t) - P_i(t-1) - R_{i,\mathrm{up}},\ 0\} + \lambda_{i,\mathrm{MT}}^{c} \min\{P_i(t) - P_i(t-1) + R_{i,\mathrm{down}},\ 0\}, \tag{47}$$

$$C_{\mathrm{ES}}^{S}(t) = -\lambda_{\mathrm{ES}}^{S} \cdot S_{es} \max\{SOC(t) - SOC_{\mathrm{max}},\ 0\} + \lambda_{\mathrm{ES}}^{S} \cdot S_{es} \min\{SOC(t) - SOC_{\mathrm{min}},\ 0\}, \tag{48}$$

where $C_{i,\mathrm{MT}}^{c}$ represents the penalty for the $i-$th distributed MT exceeding the ramping constraint; $\lambda_{i,\mathrm{MT}}^{c}$ indicates the penalty coefficient for the ramp constraint penalty of the $i-$th MT; $C_{\mathrm{ES}}^{S}(t)$ represents the penalty for energy storage exceeding the SOC constraint; and $\lambda_{\mathrm{ES}}^{S}$ denotes the penalty coefficient for the SOC constraint penalty.

For ES, the SOC at the end of the current scheduling cycle is the same as the SOC at the beginning of the next scheduling cycle. To ensure that the SOC at the end of the current cycle does not affect the scheduling ability of ES in the next scheduling cycle, we aim for the SOC at the end and beginning of each scheduling cycle to be as equal as possible. Therefore, we have designed an exponential reset penalty for the SOC of the energy storage as

$$C_{\mathrm{ES}}^{r} = \lambda_{\mathrm{ES}}^{r} \cdot (e^{\delta \cdot t} - 1) \cdot [SOC(t) - SOC(0)]^2, \tag{49}$$

where C_{ES}^{r} denotes the reset penalty for energy storage; $\lambda_{\mathrm{ES}}^{r}$ indicates the penalty coefficient of the reset penalty; δ symbolizes the exponential coefficient of the reset penalty. Equation (49) indicates that the reset penalty for energy storage is small at the beginning of a scheduling cycle and increases over time. It reaches its maximum at the end of the scheduling cycle.

In summary, the upper layer optimal scheduling aims to minimize the operating costs of the islanded microgrid during a scheduling cycle by reasonably controlling the power of distributed MTs and energy storage. Therefore, the total reward function of the deep reinforcement learning agent for the islanded microgrid can be expressed as

$$R = -\sum_{t=1}^{T} [F(t) + \sum_{i=1}^{n} (C_{\mathrm{ES}}^{S}(t) + C_{\mathrm{ES}}^{r}(t))], \tag{50}$$

where R indicates the total reward of the deep reinforcement learning agent for the islanded microgrid.

3.2.2. Twin Delayed Deep Deterministic Policy Gradient Algorithm

The twin delayed deep deterministic policy gradient (TD3) algorithm is an improved version of the DDPG algorithm, which primarily addresses the overestimation problem of DDPG and improves its convergence speed [30]. The TD3 algorithm contains six neural networks: an actor network, two critic networks (critic1 and critic2), and their corresponding target networks. The update process of these six neural networks is depicted in Figure 3.

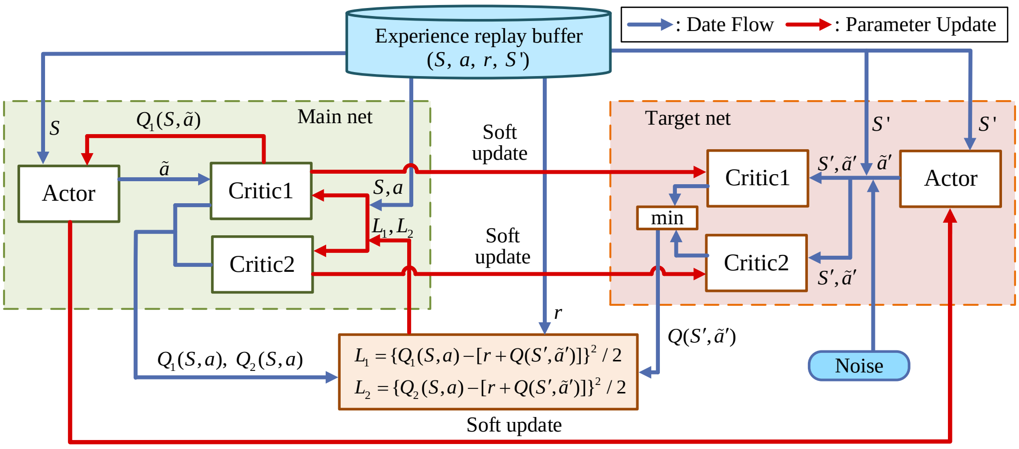


Figure 3. Update process of six neural networks in TD3 algorithm.

First, the agent obtains a tuple $(S,\ a,\ r,\ S')$ from the experience replay buffer by random sampling. The actor network then outputs an action signal $\widetilde{a}$ based on the state S in the tuple. Subsequently, the critic1 network generates the evaluation $Q(S,\widetilde{a})$ for the action under the state S. Finally, the actor network updates its network parameters θ using gradient ascent to maximize $Q(S,\widetilde{a})$. To address the overestimation problem that often occurs during the iterative update process, TD3 uses target networks and double Q−learning to improve the above process. In other words, the TD3 algorithm uses two critic networks and two target critic networks to enhance the learning effect of the algorithm. The critic networks are trained with the action and state saved in the buffer, while the next state and the predicted action by the target actor network train the target critic networks. Specifically, the critic1 and critic2 networks generate evaluation $Q_1(S,a)$ and $Q_2(S,a)$, respectively, for the action a in the tuple. The target actor network outputs an action $\widetilde{a}'$ based on the state S'. Following that, the target critic1 and critic2 networks generate evaluation $Q_1(S',\widetilde{a}')$ and $Q_2(S',\widetilde{a}')$, respectively, for the action $\widetilde{a}'$ under the state S'. The minimum of $Q_1(S',\widetilde{a}')$ and $Q_2(S',\widetilde{a}')$ is considered as $Q(S',\widetilde{a}')$. The loss functions L_1 and L_2 can be calculated using [31]:

$$L_1 = \frac{1}{2}\{Q_1(S,a) - [r + Q(S',\widetilde{a}')]\}^2 \tag{51}$$

$$L_2 = \frac{1}{2}\{Q_2(S,a) - [r + Q(S',\widetilde{a}')]\}^2 \tag{52}$$

Finally, the critic1 and critic2 networks' neural network parameters ω_1 and ω_2 are updated using gradient descent to minimize L_1 and L_2, respectively. Finally, the target actor network, target critic1 network, and target critic2 network are updated using soft updates. The update formula for soft updates can be given as [31]:

$$\theta'(t+1) = \theta(t) + (1-\tau)\theta'(t) \tag{53}$$

$$\omega_1'(t+1) = \omega_1(t) + (1-\tau)\omega_1'(t) \tag{54}$$

$$\omega_2'(t+1) = \omega_2(t) + (1-\tau)\omega_2'(t) \tag{55}$$

where θ_T represents the neural network parameters of the target actor network; ω_T symbolizes the neural network parameters of the target critic network; and $\tau \ll 1$ is the target smoothing factor, which indicates the update speed of the target neural network parameters.

In addition to the above−mentioned improvement methods, the TD3 algorithm also introduces noise signals following a truncated normal distribution into the output action signals of the target actor network. The truncated normal distribution can be denoted as $CN(0, \sigma^2, -z, z)$, which indicates that the variable follows a normal distribution with zero mean and variance σ^2, but the probability of the variable falling outside of $[-z, z]$ is zero. Using this probability distribution can prevent the output results of deep reinforcement learning agents from diverging because of the excessively large noise. Moreover, the TD3 algorithm adopts a delayed update strategy for the actor network and target network, i.e., updating the critic network once per episode, but updating the actor network and three target networks at every φ episode, where φ is an integer greater than or equal to one.

3.2.3. MATD3 Method for Optimal Scheduling of the Island Microgrid

To achieve accurate control, a single−agent reinforcement learning method requires a single agent to collect the state information of both distributed MTs and ES. This undoubtedly increases the communication costs of the island microgrid and operational risks in harsh climate environments, where communication line failures are common. The centralized training and distributed execution of multi−agent reinforcement learning architectures are popular as they do not require communication between agents during decision−making and only require local observation information for real−time decision−making. In this paper, the TD3 algorithm is extended to the MATD3 algorithm, which is suitable for island microgrids. Its feature is that the state and action information of both the MT and ES agents are collected and centralized into the experience replay buffer during the training process, i.e., $S = (S_{\text{MT}},\ S_{\text{ES}}), a = (a_{\text{MT}},\ a_{\text{ES}})$. We define the state information of all agents obtained by agents from the experience replay buffer as global state information and the state information obtained by agents interacting with the environment as local state information. The actor networks of both agents are updated based on their local state information, while the critic network is updated based on the global state information. Considering the MT agent as an example, its actor network outputs actions $\widetilde{a}_{\text{MT}}$ based on its local state S_{MT}. The critic network of the MT agent generates evaluations $Q(S,\ \widetilde{a}_{\text{MT}})$ and $Q(S,\ a_{\text{MT}})$ for $\widetilde{a}_{\text{MT}}$ and a_{MT}, respectively, based on the global state information S obtained from both agents in the experience replay buffer. Finally, the actor network of the MT agent updates its network parameters θ based on $Q(S,\ \widetilde{a}_{\text{MT}})$, while the critic network updates its network parameters ω based on $Q(S,\ a_{\text{MT}})$. The same process is applied to the ES agent.

To simplify the notation, let $v \in \{\text{MT, ES}\}$. Each episode of the MATD3 agent's training contains T time steps, and the training process repeats M times to ensure that the algorithm converges. The specific working process of the proposed MATD3 algorithm for island microgrids is detailed in Algorithm 1.

Algorithm 1: MATD3−Based Optimized Scheduling Method for Island Microgrids

```
Initialize $\tau$, $\theta_v$, $\omega_{v1}$, $\omega_{v2}$ and experience replay buffer D
for episode = 1 to M do
    Initialize random process N for action exploration
    for $t$ = 1 to $T$ do
        The MT agent and the ES agent observe their respective state spaces $S_{\mathrm{MT}}(t)$ and $S_{\mathrm{ES}}(t)$ from their own environments
        Choose the power actions $a_{\mathrm{MT}}$ and $a_{\mathrm{ES}}$ of distributed MTs and energy storage, respectively
        The island microgrid operates according to actions $a_{\mathrm{MT}}$ and $a_{\mathrm{ES}}$, and it gets the real island microgrid environmental reward $r(t)$ via (50)
        The MT agent and the ES agent observe the new state spaces $S'_{\mathrm{MT}}(t)$ and $S'_{\mathrm{ES}}(t)$ from the island microgrid environment, respectively
        Merge $(S_{\mathrm{MT}},\ a_{\mathrm{MT}},\ r,\ S'_{\mathrm{MT}})$ and $(S_{\mathrm{ES}},\ a_{\mathrm{ES}},\ r,\ S'_{\mathrm{ES}})$ into $(S,\ a,\ r,\ S')$ and store $(S,\ a,\ r,\ S')$ in D
        $S_{\mathrm{MT}} \leftarrow S'_{\mathrm{MT}},\ \ S_{\mathrm{ES}} \leftarrow S'_{\mathrm{ES}}$
        for MT agent and ES agent do
            Sample a random mini−batch of size $H$ $(S,\ a,\ r,\ S')$ from D
            Update critic network parameters $\omega_{v1}$ and $\omega_{v2}$ by minimizing the loss $L_{v1}$ and $L_{v2}$, respectively
            Update actor parameter $\theta_v$ every two critic updates by maximizing $Q_v(S,\widetilde{a}_v)$
        end
        Update target network parameters for MT agent and ES agent via (53)−(55)
    end
end
```

3.2.4. Two−Stage Deep Reinforcement Learning Agent Training Method

Owing to the various complex constraints in the island microgrid, deep reinforcement learning agents often require considerable time to compute the optimum strategy without prior knowledge. Furthermore, these agents are prone to get stuck in local optimal solutions, or output actions that converge to bounds [32]. This phenomenon can be attributed to the utilization of a neural network within the TD3 algorithm for approximating the action value function in reinforcement learning [33]. During the initial stages of agent training, the approximation error associated with this neural network is considerably high. When this error surpasses a certain threshold, the agent becomes incapable of effectively mitigating it over numerous training iterations. As a result, the algorithm fails to accurately estimate the value of the agent's output actions, thereby introducing a bias towards specific action choices. Moreover, despite the TD3 algorithm incorporating several measures to address the problem of over− and underestimation, this issue persists, particularly when the initial random seeds deviate significantly from the optimal policy [34]. If any of the critic networks in the TD3 algorithm overestimate the value for certain state−action pairs, the algorithm becomes inclined to select these actions disproportionately. Conversely, if a critic network underestimates the value of specific actions, they are more likely to be disregarded, consequently limiting the exploration space of the algorithm. In addition, since the TD3 algorithm is a deterministic policy method, it is sensitive to the initial random seed in the experience replay buffer [35]. Different initial random seeds directly affect the learning effect of the TD3 agent. Imitation learning has the advantages of fast training speed and good convergence. Pre−training the deep reinforcement learning agent through imitation learning can initialize the experience replay buffer and neural network parameters of the agent, thereby considerably improving the training speed and learning efficiency of the agent. Therefore, this study adopted a two−stage training method by combining the imitation learning pre−training stage and the reinforcement learning training stage, which means that the agent is initially trained using pre−provided data, and subsequently, the training continues by modifying the reward function to further explore different policies. The working process is shown in Figure 4.

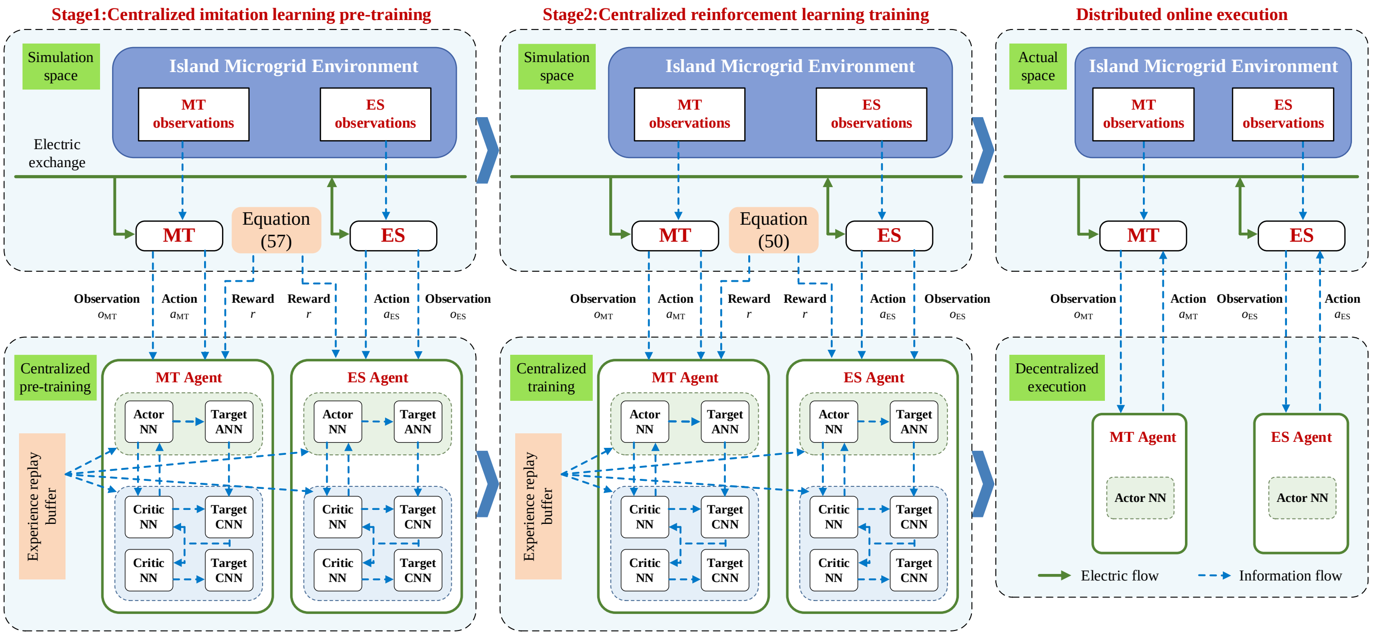


Figure 4. Two−stage deep reinforcement learning training method.

- Stage 1: Imitation learning pre−training stage

The expert scheduling strategy for the day is given by the scheduling personnel of the island microgrid, which can be described as

$$D = \{(P'_{\mathrm{MT}}(1),\ P'_{\mathrm{ES}}(1)),\ (P'_{\mathrm{MT}}(2),\ P'_{\mathrm{ES}}(2)),\ \cdots,\ (P'_{\mathrm{MT}}(T),\ P'_{\mathrm{ES}}(T))\}, \tag{56}$$

where $P'_{\mathrm{MT}}(i)$ represents the total output power suggested by the scheduling personnel for the i−th time step for MTs, and $P'_{\mathrm{ES}}(i)$ represents the charging/discharging power suggested by the scheduling personnel for the i−th time step for ES. Imitation learning aims to minimize the difference between the actions taken by the MT agent and the ES agent and those suggested by the scheduling personnel in each time step. Therefore, the reward function for the imitation learning pre−training stage is defined as

$$R' = -K'\sum_{i=1}^{T}[(P'_{\mathrm{MT}}(i) - P^{m}_{ref}(i))^2 + (P'_{\mathrm{ES}}(i) - P_{\mathrm{ES},\ ref}(i))^2], \tag{57}$$

where $P^{m}_{ref}(i)$ represents the reference signal of the total output power of the MTs output by the MT agent at the i−th time step, and $P_{\mathrm{ES},\ ref}(i)$ represents the reference signal of the ES charging/discharging power output by the ES agent at the i−th time step. Subsequently, the MT agent and the ES agent will use the method described in Section 3.2.3 to pretrain the agents with the objective of maximizing the reward R'. As the reward function in (57) is a simple quadratic function with a unique extremum, the convergence speed of the agents' iterative calculations during the imitation learning training phase will be much faster than that of reinforcement learning.

- Stage 2: Reinforcement learning training stage

Upon completing the pre−training of imitation learning, we obtained the pre−trained MT and ES agents. Subsequently, we modified the reward function to (50). The MT and ES agents aim to maximize the real environment reward R and continue training based on the pre−trained MT and ES agents, which is performed to search for the optimum strategy for island microgrid scheduling.

The two−stage training method described above was conducted in an offline simulation environment. Upon completing the proposed two−stage training, both the MT and ES agents can utilize their respective actor networks to make distributed online decisions for the island microgrid in a real−world environment.

In summary, the working flowchart of the dual−layer distributed optimal operation method proposed in this section is shown in Figure 5.

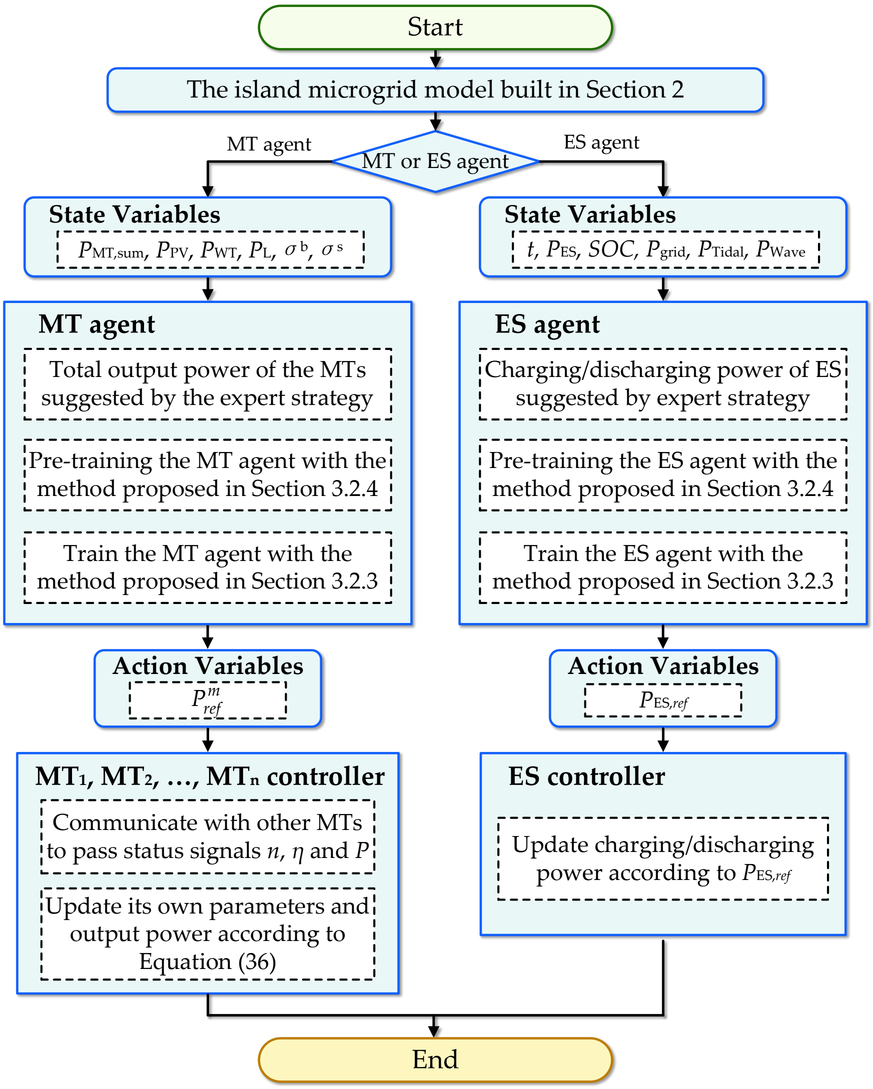

Start
The island microgrid model built in Section 2
MT agent
MT or ES agent
ES agent
State Variables
$P_{\mathrm{MT,sum}}$, P_{PV}, P_{WT}, P_{L}, σ^{b}, σ^{s}
State Variables
t, P_{ES}, SOC, P_{grid}, P_{Tidal}, P_{Wave}
MT agent
Total output power of the MTs suggested by the expert strategy
Pre-training the MT agent with the method proposed in Section 3.2.4
Train the MT agent with the method proposed in Section 3.2.3
ES agent
Charging/discharging power of ES suggested by expert strategy
Pre-training the ES agent with the method proposed in Section 3.2.4
Train the ES agent with the method proposed in Section 3.2.3
Action Variables
P_{ref}^{m}
Action Variables
$P_{\mathrm{ES},ref}$
MT₁, MT₂, …, MTₙ controller
Communicate with other MTs to pass status signals n, η and P
Update its own parameters and output power according to Equation (36)
ES controller
Update charging/discharging power according to $P_{\mathrm{ES},ref}$
End

Figure 5. Working flowchart of the dual−layer distributed optimal operation method.

4. Numerical Simulation Analysis

This section designs an island microgrid system with PV power generation, wind power generation, tidal energy generation, wave energy generation, ES, and three distributed MTs. The neighboring MTs exchange status information via communication links. The data on PV power generation, wind power generation, tidal energy generation, wave energy generation, and load in the island microgrid are shown in Figure 6 [36]. The time−of−use pricing for electricity purchase and sale of the island microgrid from the external grid are listed in Table 2 [37]. The relevant parameters of the devices in the island microgrid are listed in Table 3. The upper and lower boundaries of each variable in the state space and action space of the MT agent and ES agent are shown in Table 4.

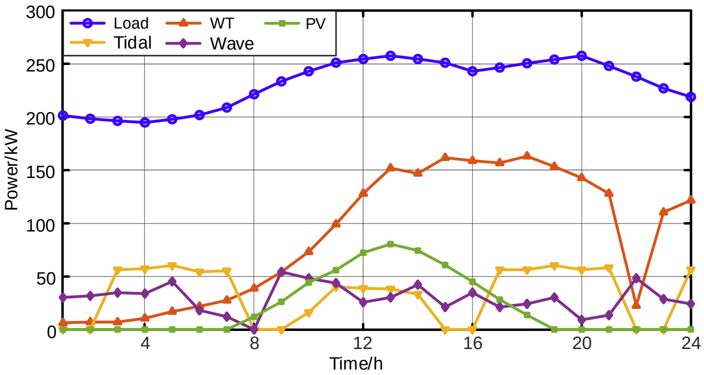


Figure 6. Power output of different types of renewable energy in the designed island microgrid.

Table 2. Time−of−use pricing for electricity purchase and sale of the island microgrid from the external grid.

Time/h	Electricity Purchase Price [CNY/(kW·h)]	Electricity Sales Price [CNY/(kW·h)]
1–6, 22–24	0.37	0.28
7–9, 14–17, 20, 21	0.82	0.65
10–13, 18, 19	1.36	0.78

Table 3. The relevant parameters of the devices in the island microgrid.

Main Parameters	Values	Main Parameters	Values
P_{nom}	200 kW	$R_{2\text{up}}$	60 kW
$P_{1\text{min}}$	0 kW	$R_{3\text{up}}$	50 kW
$P_{2\text{min}}$	0 kW	P_{chmax}	100 kW
$P_{3\text{min}}$	0 kW	P_{dismax}	100 kW
$P_{1\text{max}}$	160 kW	S_{es}	1500 kW·h
$P_{2\text{max}}$	120 kW	$SOC\ (0)$	0.5
$P_{3\text{max}}$	100 kW	α	0.0013
$R_{1\text{down}}$	80 kW	β	0.553
$R_{2\text{down}}$	60 kW	c	14.17
$R_{3\text{down}}$	50 kW	g_E	0.5
$R_{1\text{up}}$	80 kW	g_P	10

Table 4. The upper and lower boundaries of each variable in the state space and action space.

Variables	Agent	Space	Lower Boundary	Upper Boundary
$P_{\text{MT,sum}}$	MT	State	0 kW	380 kW
P_{PV}	MT	State	0 kW	200 kW
P_{WT}	MT	State	0 kW	300 kW
P_{L}	MT	State	0 kW	500 kW
σ^{b}	MT	State	0 CNY	2 CNY
σ^{s}	MT	State	0 CNY	2 CNY
t	ES	State	0:00	24:00
P_{ES}	ES	State	−100 kW	100 kW
SOC	ES	State	0	1
P_{grid}	ES	State	−1000 kW	1000 kW
P_{Tidal}	ES	State	0 kW	200 kW
P_{Wave}	ES	State	0 kW	200 kW
P_{ref}^{m}	MT	Action	0 kW	380 kW
$P_{\text{ES},ref}$	ES	Action	−100 kW	100 kW

4.1. Simulation Analysis of Lower Layer Adaptive Consensus Control

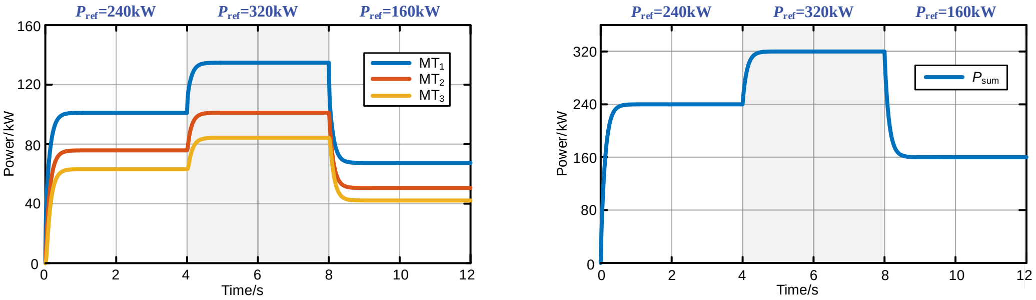
Assuming that all MTs in the island microgrid are schedulable, the controller of MT_1 can receive reference signals provided by the upper layer. The adjacency matrices of the three state signals in the MTs communication network graph are A_n = A_η = [0, 500, 500; 500, 0, 500; 500, 500, 0] and A_P = [0, 20, 20; 20, 0, 20; 20, 20, 0], and the leading adjacency matrix of the state signal P is diag{20, 0, 0}. The unit of the reference signal for the total output power of the MTs is MW, ranging from [0, 0.38]. As listed in Table 3, the capacity ratio of the three MT is 0.8:0.6:0.5, so $\eta_1(0)$ = 0.8, $\eta_2(0)$ = 0.6, and $\eta_3(0)$ = 0.5. It can be calculated that 1.205 < $\eta^{\mathrm{T}}\eta$ < 1.25. It can be determined that L_n, L_η, L_P, B, and P_{ref} can satisfy the requirements of (41). We design two scenarios for the lower layer control of the island microgrid.

Scenario 1: The reference signal provided by the upper layer changes without MT plugging in or out;

Scenario 2: The reference signal provided by the upper layer remains unchanged with MT plugging in or out.

4.1.1. Assessment of Control Performance when the Reference Signal Changes

In this section, we assess the control performance of the lower layer control for Scenario 1. The upper layer provides an initial power reference signal P_{ref} of 240 kW. At t = 4 s, P_{ref} changes to 320 kW, and at t = 8 s, P_{ref} changes to 160 kW. The changes in the output power of each of the three MTs and the total output power within a period of 12 s are shown in Figure 7.



(**a**) (**b**)

Figure 7. Distributed MT output power changes when the upper layer reference signal changes. (**a**) Output power of the three MTs; (**b**) total output power of the three MTs.

As shown in Figure 7a,b, when three MTs are connected at t = 0 s, the total output power of the three MTs can rapidly converge to the reference value and achieve accurate power allocation according to a capacity ratio of 0.8:0.6:0.5. Even when the reference signal provided by the upper layer changes at t = 4 s and t = 8 s, the three MTs can still achieve a total output power equal to the reference signal, while maintaining power allocation according to their respective capacities.

4.1.2. Plug−and−Play Performance Assessment of Distributed MTs

Here, we assess the control performance of the lower layer control in Scenario two. The power signal $P_{ref} = 240$kW provided by the upper layer remains constant. At t = 4 s, MT_3 is plugged out from the island microgrid due to a fault. At t = 8 s, MT_3 is repaired and plugged back. To assess the plug and play performance of the lower layer control scheme, we conducted experiments under two conditions: without adding the plug−and−play control method, and with its addition, as described in Section 3.1.3. We recorded the changes in the output power of the three MTs and the total output power using Figures 8 and 9 for the first and second conditions, respectively.

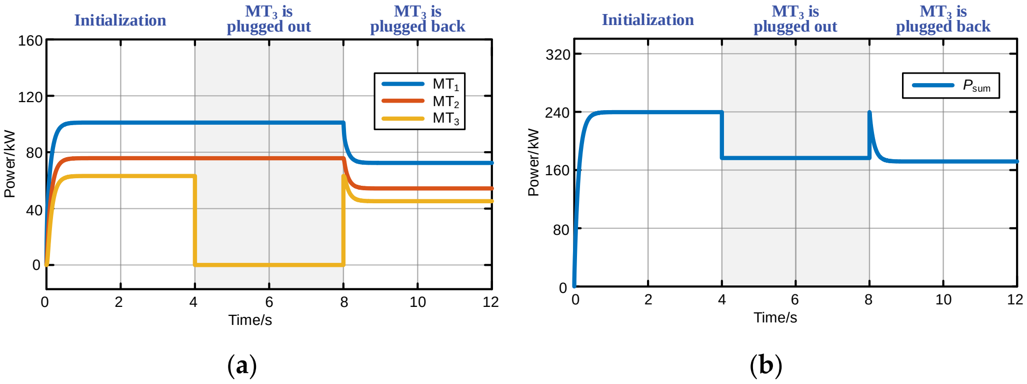


Figure 8. Without adding the plug−and−play control method, the output power changes of distributed MTs when MT_3 is being plugged in and back. (**a**) Output power of the three MTs; (**b**) total output power of three MTs.

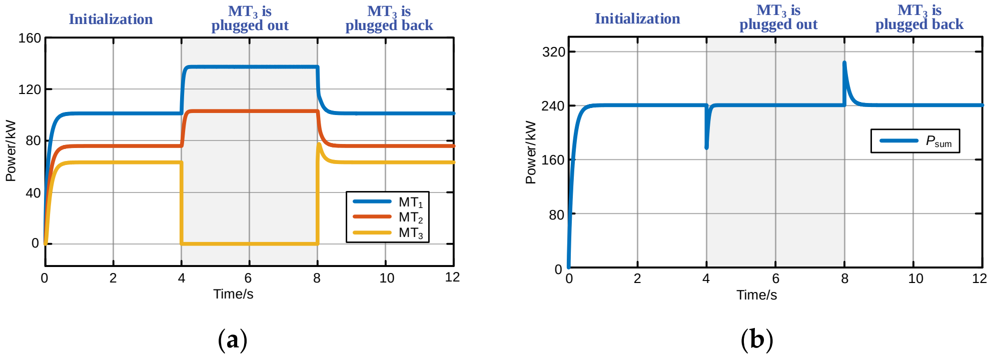


Figure 9. Adding the plug−and−play control method, the output power changes of distributed MTs when MT_3 is being plugged in and back. (**a**) Output power of the three MTs; (**b**) total output power of three MTs.

As shown in Figure 8a, when using the adaptive consensus control method proposed in this paper without the plug−and−play control, the output power of the MTs can be allocated according to their capacity. However, when MT_3 is plugged out at t = 4 s, as shown in Figure 8b, the output powers of MT_1 and MT_2 remain unchanged and the total output power of the MTs cannot track the power reference signal provided by the upper layer. As depicted in Figure 9a,b, by adding the plug−and−play control method proposed in this paper, MT_1 and MT_2 can quickly adjust their output power after MT3 is plugged out at t = 4 s, so that the sum of their output powers can track the power reference signal. Moreover, the output power of MT_1 and MT_2 can be accurately allocated according to their capacity. When MT_3 is plugged back at t = 8 s, MT_1, MT_2, and MT_3 can adaptively adjust their output power so that the total output power of the three MTs can still track the power reference signal while achieving capacity−based power allocation.

In summary, the lower layer control method proposed in this paper makes the total power of MTs follow the upper layer reference signal while ensuring that the distributed MT output power is distributed according to capacity. Furthermore, the method can achieve the plug−and−play capability of distributed MTs.

4.2. Simulation Analysis of Upper Layer Optimization Scheduling

4.2.1. Analysis of Simulation Results of Upper Layer Optimal Scheduling

As assumed, the expert strategy provided by the scheduling personnel based on the daily environment and load information is shown in Figure 10.

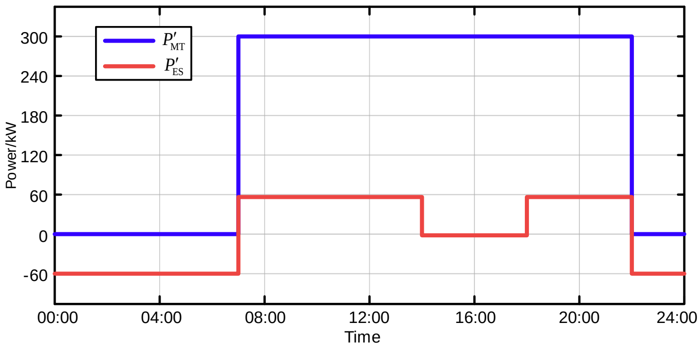


Figure 10. Expert strategy provided by the scheduling personnel.

Based on Figure 10 and Table 2, which provide detailed information about the power grid's electricity prices and time periods, we can establish a more comprehensive understanding of the expert strategy for optimizing the operation of the island microgrid.

The expert strategy identifies two distinct low electricity price periods during the day. The first period spans from 00:00 to 07:00, and the second period spans from 22:00 to 24:00. These periods are considered ideal for cost−saving measures. In these time slots, the strategy recommends terminating the operation of MTs and instead prioritizes charging the ES at a power rate of 60 kW. By doing so, excess electricity available during these low−price periods can be efficiently stored for later use. Conversely, the power grid experiences high electricity price periods and medium electricity price periods between 07:00 and 22:00, which indicate the peak demand periods. Within these periods, the expert strategy suggests a total output power of 300 kW for the MTs. This higher power output ensures that the microgrid can meet the electricity demands during these peak hours. To further optimize the system, the expert strategy proposes specific actions during different time intervals within the high electricity price periods. From 07:00 to 14:00 and again from 18:00 to 22:00, the expert strategy advises discharging the energy storage system at a rate of 60 kW. By utilizing the stored energy during these specific time frames, the microgrid can reduce its reliance on the MTs and therefore minimize costs. During the medium electricity price period of 14:00 to 18:00, the expert strategy recommends that the charging and discharging power of the energy storage system be set to zero. This decision likely reflects the moderate electricity prices during this period, indicating that there is no significant advantage in either charging or discharging the ES system.

Although the expert strategy provides a framework for optimizing the operation of the island microgrid, it is acknowledged that it may not be highly precise. To address this limitation, we adopted the two−stage MATD3 algorithm, which will be used to compute the optimal scheduling strategy for the microgrid. This algorithm takes into account various factors, such as electricity prices, load demands, and storage capacity, to determine the most efficient operation schedule for the microgrid, surpassing the limitations of the initial expert strategy.

According to the upper layer optimal scheduling method for island microgrids proposed in this paper, the MT agent and ES agent are first pre−trained for 250 episodes using the expert policy illustrated in Figure 10 and imitation learning method. Subsequently, we switched to the reward function in the real environment and trained for 500 episodes.

Figures 11 and 12 represent the reward curve and constraint penalty curve, respectively. According to Figure 11, after 400 episodes, the reward curve has basically converged, indicating that the agent has found an optimum scheduling strategy for the island microgrid. As illustrated in Figure 12, in the later stage of reinforcement learning training, the constraint penalty curve remains at zero, indicating that the scheduling strategy will not exceed the operational constraints of the island microgrid, and the island microgrid can operate safely.

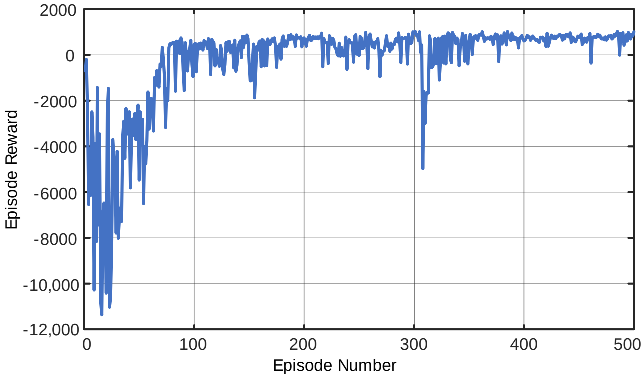


Figure 11. Reward curve of the proposed algorithm.

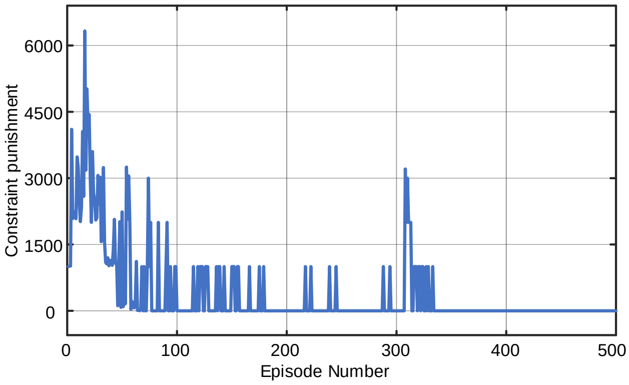
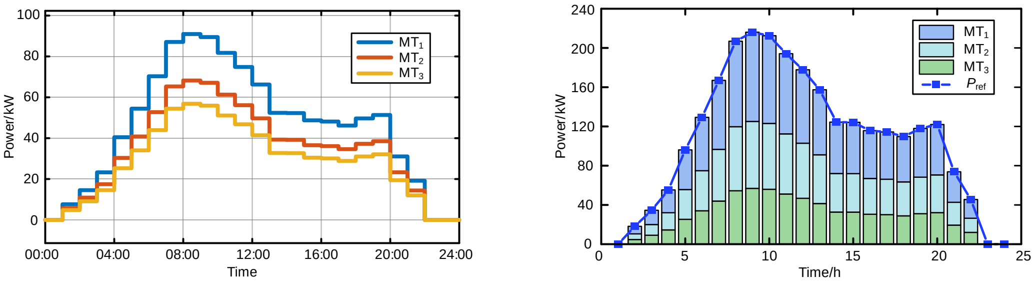


Figure 12. Constraint penalty curve of the proposed algorithm.

As depicted in Figure 13a, MT$_1$, MT$_2$, and MT$_3$ can strictly allocate their output power in a ratio of 8:6:5 during the 24−h scheduling period, and owing to the effect of the ramp constraint penalty of (47), none of them exceed their respective ramp constraint. Meanwhile, as shown in Figure 13b, the total output power of the three MTs can track the reference signal provided by the upper layer. As shown in Figure 14a,b, owing to the effect of constraints penalty (48) and (49) in the reward function, the SOC of ES during the 24−h scheduling period did not exceed the range of the constraints. Moreover, the SOC of ES at the end of the scheduling period could return to a position that is very close to the SOC at the beginning of the scheduling period.



(**a**) (**b**)

Figure 13. Distributed MT output power within a 24−h scheduling cycle. (**a**) Output power of the three MTs; (**b**) stacked bar chart of the output power of three MTs and the upper reference signal.

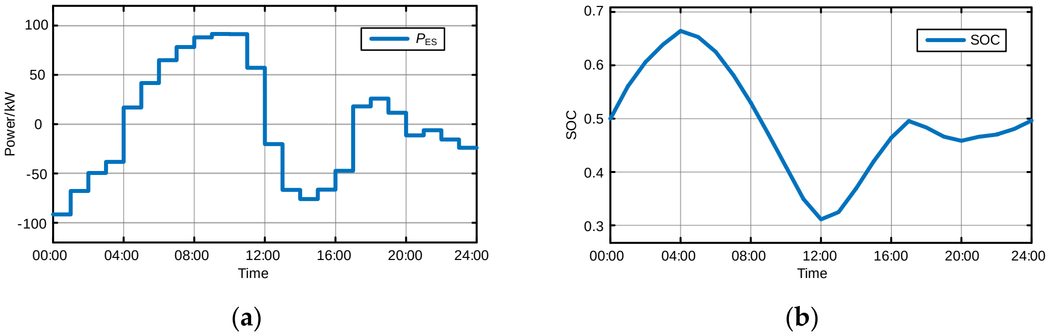


Figure 14. (**a**) Charging/discharging power of energy storage in the 24−h scheduling period; (**b**) SOC of ES within a 24−h scheduling period.

The power of various energy sources in the island microgrid during the 24−h scheduling period is shown in Figure 15. From hours 1–4, the renewable energy output is low, and the electricity price in the grid is also low. Therefore, the island microgrid purchases electricity from the external grid to maintain the supply−demand balance. Concurrently, the energy storage system charges using the electricity the grid provides. From hours 5–6, in order not to exceed the ramp constraints of distributed MTs, MTs slowly increase their output power. As the load is low during this time, the island microgrid sells a small amount of excess electricity to the external grid. From hours 7–13, the purchasing and selling price of electricity from the external grid is high, and the renewable energy output in the island microgrid begins to increase. During this period, distributed MTs continue to increase their output power, and simultaneously, the ES discharges. The island microgrid sells electricity to the external grid for higher revenue. At hour 11, the power sold to the external grid maximizes. From hours 14–17, the renewable energy output power of the island microgrid is high, and the price for selling to the external grid becomes relatively low. During this period, distributed MTs decrease their output power, and the ES charges use abundant renewable energy. Simultaneously, the island microgrid reduces the power sold to the external grid. From hours 18–20, the price for selling to the external grid becomes higher again. During this period, the output power of distributed MTs increases again, and the power sold to the external grid also increases. From hours 21–24, the electricity price in the grid is low and the renewable energy output is also low. During this period, the output power of distributed MTs decreases, and from hours 23–24, the distributed MTs stop working. The island microgrid uses the electricity purchased from the external grid and the renewable energy output to supply power to the load in the island microgrid and charge the energy storage.

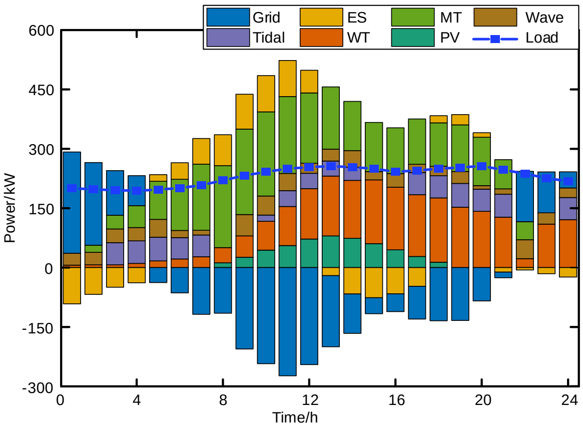
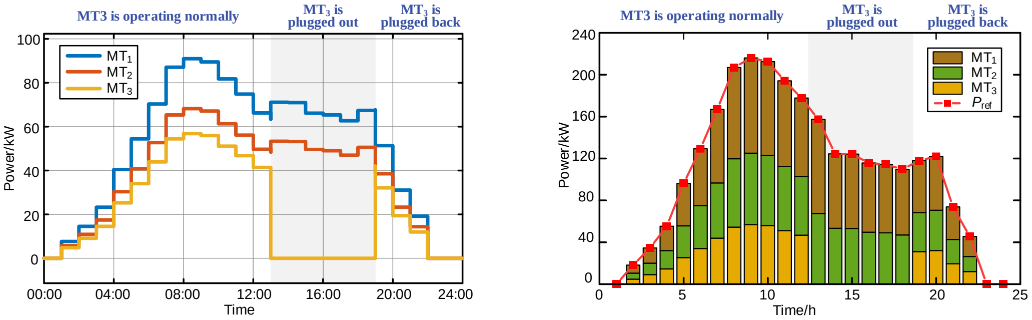


Figure 15. Load and different energy power in the 24−h scheduling period of the island microgrid.

4.2.2. Assessment of Plug—and—Play Control Performance of MTs during Optimal Scheduling

To test the plug—and—play capability of distributed MTs throughout the 24—h scheduling period, we assume that MT$_3$ is plugged out from the island microgrid at 13:00 due to a fault and is repaired and plugged back to the island microgrid at 19:00. The power output variations of the three distributed MTs and the total output power in this scenario are illustrated in Figure 16.



(**a**) (**b**)

Figure 16. Distributed MT output power within a 24—h scheduling period when MT$_3$ is being plugged in and back. (**a**) Output power of the three MTs; (**b**) stacked bar chart of the output power of three MTs and the upper reference signal.

As shown in Figure 16a,b, MT$_3$ was plugged out from the island microgrid at 13:00 due to a fault, resulting in its output power dropping to zero. At the time, the remaining operational MT$_1$ and MT$_2$ could quickly adjust their output power to maintain a ratio of 8:6 between them while ensuring that their total output power equaled the reference signal provided by the upper layer MT agent. At 19:00, MT$_3$ was repaired and plugged back into the island microgrid. At the time, the proposed lower layer adaptive consensus control method can automatically adjust the output power of the three MTs in a ratio of 8:6:5 while ensuring that their total output power is equal to the reference signal provided by the upper layer MT agent.

4.2.3. Performance Assessment of the Two—Stage Deep Reinforcement Learning Agent Training Method

To assess the effectiveness of the two—stage deep reinforcement learning agent training method proposed in this paper, we compare the performance of the proposed method with that of the traditional single—stage training method for different algorithm parameters. Three deep reinforcement learning algorithm parameters are involved: actor network learning rate, critic network learning rate, and the standard deviation of the noise. Seven different parameter selection schemes are chosen in this study, as listed in Table 5. Scheme 1 is the parameter scheme used in Sections 4.2.1 and 4.2.2. In Schemes 2, 3, and 4, the actor and critic network learning rates differ, while the standard deviation of the noise is the same. Conversely, in Schemes 5, 6, and 7, the standard deviation of noise differs, while the actor and critic network learning rates are the same.

Figure 17 represents the reward curves of the traditional single—stage deep reinforcement learning training method with different algorithm parameters. Comparing the four curves of Schemes 1, 2, 3, and 4, it is evident that when the standard deviation of the noise is fixed, the learning rates of the actor and critic networks are 1×10^{-2}, 1×10^{-2} or 5×10^{-3}, 8×10^{-3}, respectively, and the agent can eventually find a relatively satisfactory scheduling policy. However, the convergence rate of its reward curve is slow, and in the early stage of training, the reward curve fluctuates significantly. When the learning rate of both the actor and critic networks is 5×10^{-3}, the reward curve keeps fluctuating slightly below the optimum strategy; however, the convergence of the reward curve is fast. This suggests that the agent's reward curve is trapped in a local optimum, and it fails to escape from it until the end of the training period. If the learning rates of the actor and critic networks are 1×10^{-2} and 5×10^{-3}, respectively, then the agent's output actions converge to the boundary, resulting in a low and stable reward during the early and middle stages of training. In the later stages of training, the output action values of the agent jump out of the boundary. However, the agent is still unable to find an effective control strategy throughout the training period.

Table 5. Seven different parameter selection schemes in this paper.

Schemes	Actor Network Learning Rate	Critic Network Learning Rate	Standard Deviation of Noise
1	1×10^{-2}	1×10^{-2}	4
2	1×10^{-2}	5×10^{-3}	4
3	5×10^{-3}	5×10^{-3}	4
4	5×10^{-3}	8×10^{-3}	4
5	1×10^{-2}	1×10^{-2}	2
6	1×10^{-2}	1×10^{-2}	3
7	1×10^{-2}	1×10^{-2}	5

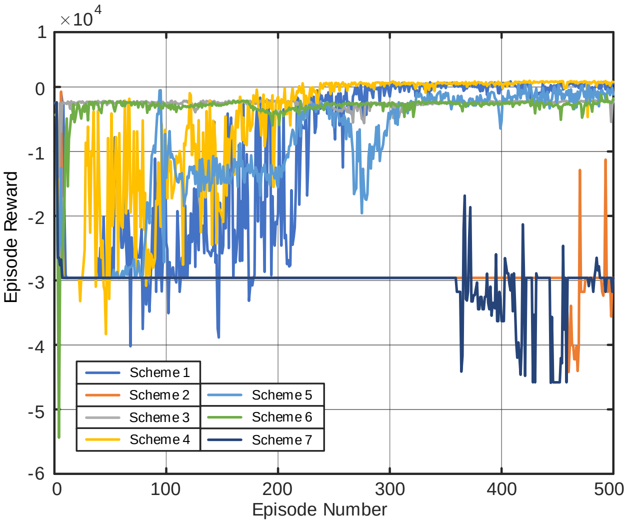


Figure 17. Reward curves of the traditional single−stage deep reinforcement learning training method with different algorithm parameters.

Similarly, by comparing the four curves of Schemes 1, 5, 6, and 7, it can be seen that when the learning rates of the critic and actor networks are fixed, the convergence rate of the agent's reward curve is slow when the standard deviation of the noise is either two or four, and the reward value remains low during the later stages of training. When the standard deviation of the noise is three, the agent's reward curve is trapped in a local optimum until the end of training. When the standard deviation of the noise is five, the agent's output actions converge to the boundary, resulting in a low and stable reward during the early and middle stages of training. Only in the later stages of training do the agent's output action values jump out of the boundary. However, the agent still fails to find an effective control strategy throughout the training period.

Based on the above analysis, we can infer that the traditional single−stage deep reinforcement learning training method is sensitive to changes in the learning rates of the actor and critic networks, as well as the standard deviation of the noise. Therefore, it can be inferred that choosing poor deep reinforcement learning algorithm parameters for the agents can result in the reward curve being trapped in a local optimum or the agent's output actions converging to the boundaries.

Figure 18 represents the reward curves of the two−stage deep reinforcement learning training method proposed in this paper with different algorithm parameters. Table 6 presents a comparison of the average reward values of the two methods in the last 10 episodes of training. As shown in Figure 18, using the two−stage training method proposed in this paper, the agent can find the optimum scheduling policy for the island microgrid in the last 100 episodes of training. The result indicates that the two−stage training method proposed in this paper has good adaptability to different algorithm parameters. As listed in Table 6, the two−stage training method proposed in this paper has a higher average reward value in the last 10 episodes, indicating that the proposed training method enables the agent to find a better scheduling strategy for the island microgrid than the traditional single−stage training method.

Table 6. Average reward values for single− and two−stage training methods in the last ten episodes.

Schemes	Two−Stage Training	Single−Stage Training
1	779.66	189.06
2	821.93	−28,863.41
3	516.11	−2556.68
4	703.45	344.69
5	552.89	−943.76
6	448.92	−2163.31
7	391.78	−30,037.94

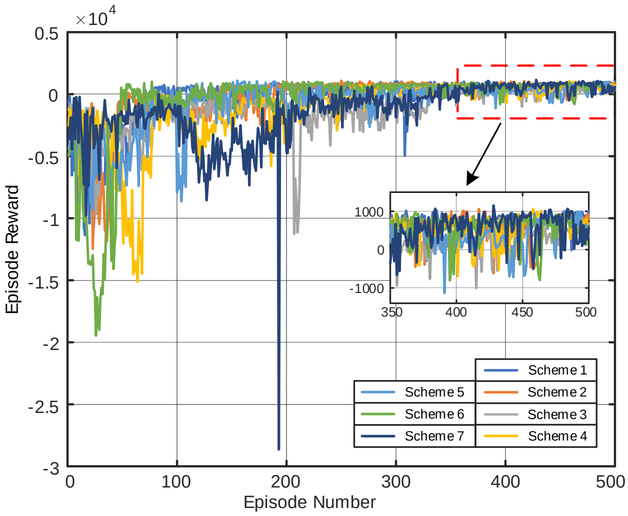


Figure 18. Reward curves of the two−stage deep reinforcement learning training method proposed in this paper with different algorithm parameters.

In summary, it can be concluded that the two−stage deep reinforcement learning training method proposed in this paper can reduce the sensitivity of the deep reinforcement learning training process to algorithm parameters, thereby reducing the tuning difficulty of the deep reinforcement learning algorithm. Furthermore, the proposed method improves the training effectiveness of the agents.

4.2.4. Comparative Analysis of Algorithms

To verify the superiority of the two−stage MATD3 algorithm proposed in this paper for the optimal scheduling of island microgrids, we compare the proposed method with other deep reinforcement learning algorithms currently applied to microgrid optimal scheduling, including the MATD3 [38], TD3 [39], MASAC [40], SAC [41], MADDPG [42], and DDPG algorithms [43]. Each algorithm was trained for 500 episodes in the island microgrid environment model established in this paper. The training results of different algorithms obtained are illustrated in Figures 19 and 20.

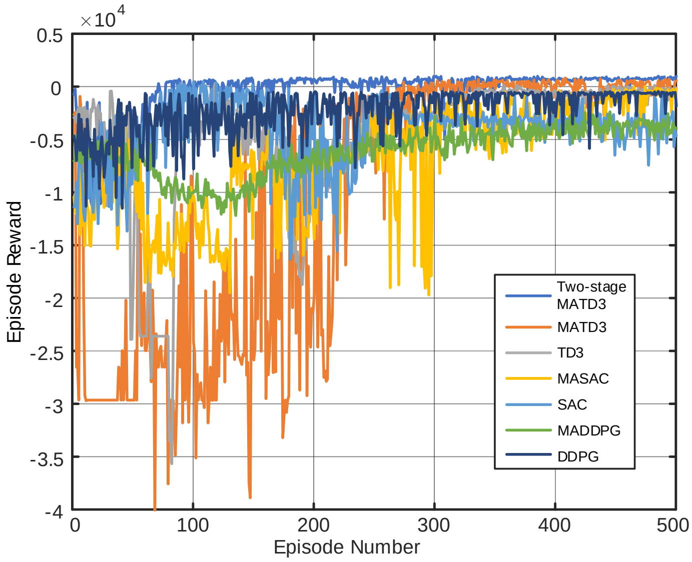


Figure 19. Reward curves for different deep reinforcement learning algorithms.

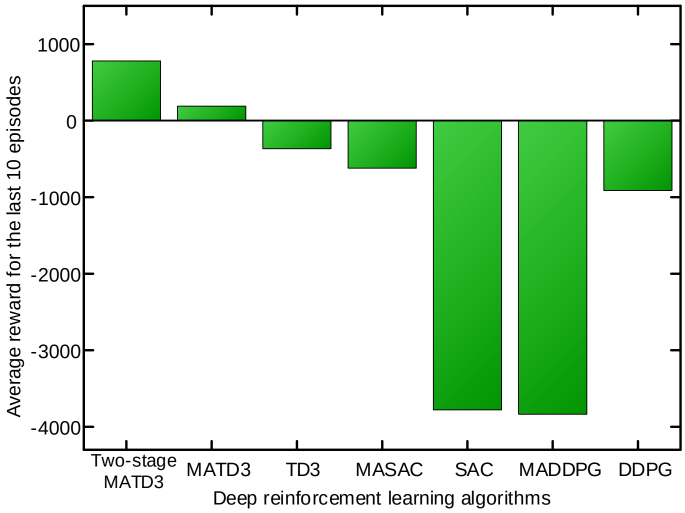


Figure 20. Average reward values of the last 10 rounds of different deep reinforcement learning algorithms.

From Figure 19, it can be seen that the algorithm proposed in this paper exhibits superior training performance compared to other deep reinforcement learning algorithms. Specifically, during the early stages of training, the proposed algorithm exhibits a faster convergence rate and can discover the optimal scheduling strategy more quickly. During the later stages of training, the proposed algorithm receives a larger reward compared to other algorithms. In addition, the reward function curve has fewer fluctuations, indicating that the algorithm has higher stability. From Figure 20, it is evident that the two−stage MATD3 algorithm proposed in this paper has a higher average reward value. It indicates that the proposed algorithm can find a better optimal scheduling strategy for island microgrids compared to other algorithms.

5. Conclusions

A two−layer distributed optimal operation method is proposed for island microgrids. The lower layer ensures the consistent operational state of distributed MTs within the microgrid. The upper layer ensures the economic operation of the microgrid. For the lower layer, a new adaptive consensus control method which ensures that the output power of the distributed MTs is allocated according to their capacity is proposed. Moreover, the proposed method ensures that the total output power of the MTs follows the reference signal provided by the upper layer, regardless of whether an MT is plugged in or out. For the upper layer, a two−stage MATD3 method is proposed; this method improves the training effectiveness of the agent and reduces its sensitivity to the deep reinforcement learning algorithm parameters. This paper provides a new solution to the problem of distributed control and optimal scheduling of island microgrids containing a high percentage and multiple types of renewable energy. It is expected that this paper will provide a useful reference for the development of future island microgrid power systems.

Author Contributions: Conceptualization, Z.Z. and B.Z.; methodology, Z.Z. and B.Z.; software, Z.Z. and G.L.; validation, G.L. and P.G.; formal analysis, B.Z. and P.G.; data curation, B.Z.; writing—original draft preparation, Z.Z.; writing—review and editing, Z.Z. and J.H.; visualization, B.L.; supervision, J.H. and B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Foundation of China, grant number U22B20115, in part by the Applied Fundamental Research Program of Liaoning Province, grant number 2023JH2/101600036, in part by the Science and Technology Projects in Liaoning Province, grant number 2022-MS-110, and in part by the Guangdong Basic and Applied Basic Research Foundation, grant number 2021A1515110778.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are included within the article.

Acknowledgments: Special thanks to the Intelligent Electrical Science and Technology Research Institute, Northeastern University (China), for providing technical support for this research.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Groppi, D.; Pfeifer, A.; Garcia, D.A.; Krajačić, G.; Duić, N. A review on energy storage and demand side management solutions in smart energy islands. *Renew. Sustain. Energy Rev.* **2021**, *135*, 110183. [CrossRef]
2. Wu, Y.; Hu, M.; Liao, M.; Liu, F.; Xu, C. Risk assessment of renewable energy−based island microgrid using the HFLTS−cloud model method. *J. Clean. Prod.* **2021**, *284*, 125362. [CrossRef]
3. Mimica, M.; De Urtasun, L.G.; Krajačić, G. A robust risk assessment method for energy planning scenarios on smart islands under the demand uncertainty. *Energy* **2022**, *240*, 122769. [CrossRef]
4. Zhao, B.; Chen, J.; Zhang, L.; Zhang, X.; Qin, R.; Lin, X. Three representative island microgrids in the East China Sea: Key technologies and experiences. *Renew. Sustain. Energy Rev.* **2018**, *96*, 262–274. [CrossRef]
5. Liu, S.; Wang, X.; Liu, P.X. Impact of communication delays on secondary frequency control in an islanded microgrid. *IEEE Trans. Ind. Electron.* **2015**, *62*, 2021–2031. [CrossRef]
6. Mahmood, H.; Michaelson, D.; Jiang, J. Reactive power sharing in islanded microgrids using adaptive voltage droop control. *IEEE Trans. Smart Grid* **2015**, *6*, 3052–3060. [CrossRef]
7. Espina, E.; Llanos, J.; Burgos−Mellado, C.; Cardenas−Dobson, R.; Martinez−Gomez, M.; Saez, D. Distributed control strategies for microgrids: An overview. *IEEE Access* **2020**, *8*, 193412–193448. [CrossRef]
8. Yue, D.; He, Z.; Dou, C. Cloud−Edge Collaboration Based Distribution Network Reconfiguration for Voltage Preventive Control. *IEEE Trans. Ind. Inf.* **2023**. [CrossRef]
9. Nguyen, T.L.; Guillo−Sansano, E.; Syed, M.H.; Nguyen, V.H.; Blair, S.M.; Reguera, L.; Tran, Q.T.; Caire, R.; Burt, G.M.; Gavriluta, C.; et al. Multi−agent system with plug and play feature for distributed secondary control in microgrid—Controller and power hardware−in−the−loop Implementation. *Energies* **2018**, *11*, 3253. [CrossRef]
10. Hosseinzadeh, M.; Schenato, L.; Garone, E. A distributed optimal power management system for microgrids with plug&play capabilities. *Adv. Control Appl. Eng. Ind. Syst.* **2021**, *3*, e65.
11. Lai, J.; Lu, X.; Yu, X.; Monti, A. Cluster−oriented distributed cooperative control for multiple AC microgrids. *IEEE Trans. Ind. Inf.* **2019**, *15*, 5906–5918. [CrossRef]
12. Lu, X.; Lai, J.; Yu, X. A novel secondary power management strategy for multiple AC microgrids with cluster−oriented two−layer cooperative framework. *IEEE Trans. Ind. Inf.* **2020**, *17*, 1483–1495. [CrossRef]

13. Gao, S.; Xiang, C.; Yu, M.; Tan, K.T.; Lee, T.H. Online optimal power scheduling of a microgrid via imitation learning. *IEEE Trans. Smart Grid* **2021**, *13*, 861–876. [CrossRef]
14. Shi, W.; Li, N.; Chu, C.C.; Gadh, R. Real−time energy management in microgrids. *IEEE Trans. Smart Grid* **2015**, *8*, 228–238. [CrossRef]
15. Paul, T.G.; Hossain, S.J.; Ghosh, S.; Mandal, P.; Kamalasadan, S. A quadratic programming based optimal power and battery dispatch for grid−connected microgrid. *IEEE Trans. Ind. Appl.* **2017**, *54*, 1793–1805. [CrossRef]
16. Tabar, V.S.; Jirdehi, M.A.; Hemmati, R. Energy management in microgrid based on the multi objective stochastic programming incorporating portable renewable energy resource as demand response option. *Energy* **2017**, *118*, 827–839. [CrossRef]
17. Abdolrasol, M.G.; Hannan, M.A.; Mohamed, A.; Amiruldin, U.A.U.; Abidin, I.B.Z.; Uddin, M.N. An optimal scheduling controller for virtual power plant and microgrid integration using the binary backtracking search algorithm. *IEEE Trans. Ind. Appl.* **2018**, *54*, 2834–2844. [CrossRef]
18. Yousif, M.; Ai, Q.; Gao, Y.; Wattoo, W.A.; Jiang, Z.; Hao, R. An optimal dispatch strategy for distributed microgrids using PSO. *CSEE J. Power Energy Syst.* **2019**, *6*, 724–734. [CrossRef]
19. Raghav, L.P.; Kumar, R.S.; Raju, D.K.; Singh, A.R. Optimal energy management of microgrids using quantum teaching learning based algorithm. *IEEE Trans. Smart Grid* **2021**, *12*, 4834–4842. [CrossRef]
20. Fan, L.; Zhang, J.; He, Y.; Liu, Y.; Hu, T.; Zhang, H. Optimal scheduling of microgrid based on deep deterministic policy gradient and transfer learning. *Energies* **2021**, *14*, 584. [CrossRef]
21. Liu, J.F.; Chen, J.L.; Wang, X.S.; Zheng, J.; Huang, Q.Y. Energy Management and Optimization of Multi−energy Grid Based on Deep Reinforcement Learning. *Power Syst. Technol.* **2020**, *44*, 3794–3803.
22. Zhao, J.; Li, F.; Mukherjee, S.; Sticht, C. Deep reinforcement learning−based model−free on−line dynamic multi−microgrid formation to enhance resilience. *IEEE Trans. Smart Grid* **2022**, *13*, 2557–2567. [CrossRef]
23. Li, T.; Yang, D.; Xie, X.; Zhang, H. Event−triggered control of nonlinear discrete−time system with unknown dynamics based on HDP (λ). *IEEE Trans. Cybern.* **2021**, *52*, 6046–6058. [CrossRef]
24. Yu, L.; Xie, W.; Xie, D.; Zou, Y.; Zhang, D.; Sun, Z.; Zhang, L.; Zhang, Y.; Jiang, T. Deep reinforcement learning for smart home energy management. *IEEE Internet Things J.* **2019**, *7*, 2751–2762. [CrossRef]
25. Ji, Y.; Wang, J.H. Online optimal scheduling of a microgrid based on deep reinforcement learning. *Control Decis.* **2022**, *37*, 1675–1684.
26. De Azevedo, R.; Cintuglu, M.H.; Ma, T.; Mohammed, O.A. Multiagent−based optimal microgrid control using fully distributed diffusion strategy. *IEEE Trans. Smart Grid* **2017**, *8*, 1997–2008. [CrossRef]
27. Zhang, J.; Pu, T.; Li, Y.; Wang, X.; Zhou, X. Multi−agent deep reinforcement learning based optimal dispatch of distributed generators. *Power Syst. Technol.* **2022**, *46*, 3496–3504.
28. Olfati−Saber, R.; Fax, J.A.; Murray, R.M. Consensus and cooperation in networked multi−agent systems. *Proc. IEEE* **2007**, *95*, 215–233. [CrossRef]
29. Woo, J.; Yu, C.; Kim, N. Deep reinforcement learning−based controller for path following of an unmanned surface vehicle. *Ocean Eng.* **2019**, *183*, 155–166. [CrossRef]
30. Zhang, F.; Li, J.; Li, Z. A TD3−based multi−agent deep reinforcement learning method in mixed cooperation−competition environment. *Neurocomputing* **2020**, *411*, 206–215. [CrossRef]
31. Fujimoto, S.; Hoof, H.; Meger, D. Addressing function approximation error in actor−critic methods. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10 July 2018.
32. Du, Y.; Wu, D. Deep reinforcement learning from demonstrations to assist service restoration in islanded microgrids. *IEEE Trans. Sustain. Energy* **2022**, *13*, 1062–1072. [CrossRef]
33. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Hassabis, D. Human−level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef]
34. Meng, L.; Gorbet, R.; Kulić, D. The effect of multi−step methods on overestimation in deep reinforcement learning. In Proceedings of the International Conference on Pattern Recognition, Milan, Italy, 10 January 2021.
35. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft actor−critic: Off−policy maximum entropy deep reinforcement learning with a stochastic actor. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10 July 2018.
36. Jiang, T.; Tang, S.; Li, X.; Zhang, R.; Chen, H.; Li, G. Resilience Boosting Strategy for Island Microgrid Clusters against Typhoons. *Proc. CSEE* **2022**, *42*, 6625–6641.
37. Zhao, P.; Wu, J.; Wang, Y.; Zhang, H. Operation optimization strategy of microgrid based on deep reinforcement learning. *Electr. Power Autom. Equip.* **2022**, *42*, 9–16.
38. Chen, T.; Bu, S.; Liu, X.; Kang, J.; Yu, F.R.; Han, Z. Peer−to−peer energy trading and energy conversion in interconnected multi−energy microgrids using multi−agent deep reinforcement learning. *IEEE Trans. Smart Grid* **2021**, *13*, 715–727. [CrossRef]
39. Liu, Y.; Qie, T.; Yu, Y.; Wang, Y.; Chau, T.K.; Zhang, X.; Manandhar, U.; Li, S.; Lu, H.H.; Fernando, T. A Novel Integral Reinforcement Learning−Based Control Method Assisted by Twin Delayed Deep Deterministic Policy Gradient for Solid Oxide Fuel Cell in DC Microgrid. *IEEE Trans. Sustain. Energy* **2022**, *14*, 688–703. [CrossRef]
40. Wu, T.; Wang, J.; Lu, X.; Du, Y. AC/DC hybrid distribution network reconfiguration with microgrid formation using multi−agent soft actor−critic. *Appl. Energy* **2022**, *307*, 118189. [CrossRef]

41. Arwa, E.O.; Folly, K.A. Reinforcement learning techniques for optimal power control in grid−connected microgrids: A comprehensive review. *IEEE Access* **2020**, *8*, 208992–209007. [CrossRef]
42. Samende, C.; Cao, J.; Fan, Z. Multi−agent deep deterministic policy gradient algorithm for peer−to−peer energy trading considering distribution network constraints. *Appl. Energy* **2022**, *317*, 119123. [CrossRef]
43. Li, Y.; Wang, R.; Yang, Z. Optimal scheduling of isolated microgrids using automated reinforcement learning−based multi−period forecasting. *IEEE Trans. Sustain. Energy* **2021**, *13*, 159–169. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.