energies

MDPI

Article

Optimal Scheduling of Microgrid Based on Deep Deterministic Policy Gradient and Transfer Learning

Luqin Fan [1], Jing Zhang [1,*], Yu He [1], Ying Liu [2], Tao Hu [3] and Heng Zhang [1]

[1] College of Electrical Engineering, Guizhou University, Guiyang 550025, China; gs.lqfan17@gzu.edu.cn (L.F.); yhe7@gzu.edu.cn (Y.H.); gs.hengzhang18@gzu.edu.cn (H.Z.)
[2] Power Grid Planning Research Center of Guizhou Power Grid Corporation, Guiyang 550002, China; liuying@im.gzwy.csg
[3] Guizhou Power Grid Corporation, Guiyang 550002, China; hutao@im.gz.csg
* Correspondence: zhangjing@gzu.edu.cn

Abstract: Microgrid has flexible composition, a complex operation mechanism, and a large amount of data while operating. However, optimization methods of microgrid scheduling do not effectively accumulate and utilize the scheduling knowledge at present. This paper puts forward a microgrid optimal scheduling method based on Deep Deterministic Policy Gradient (DDPG) and Transfer Learning (TL). This method uses Reinforcement Learning (RL) to learn the scheduling strategy and accumulates the corresponding scheduling knowledge. Meanwhile, the DDPG model is introduced to extend the microgrid scheduling strategy action from the discrete action space to the continuous action space. On this basis, this paper holds that a microgrid optimal scheduling TL algorithm on the strength of the actual supply and demand similarity is proposed with a purpose of making use of the existing scheduling knowledge effectively. The simulation results indicate that this paper can provide optimal scheduling strategy for microgrid with complex operation mechanism flexibly and efficiently through the effective accumulation of scheduling knowledge and the utilization of scheduling knowledge through TL.

Keywords: microgrid; optimal scheduling; reinforcement learning; transfer learning

Citation: 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. https://doi.org/10.3390/en14030584

Received: 7 December 2020
Accepted: 15 January 2021
Published: 23 January 2021

1. Introduction

Microgrid is a small-scale power grid, composed of distributed power generation, load, energy storage devices, and energy conversion devices, which can effectively improve the stability and power quality of a large number of distributed power sources connected to the main grid, and realize the flexible application of distributed power generation [1]. However, the intermittence and instability of distributed generation make energy management more difficult. How to manage the energy of microgrid efficiently is a challenge for microgrid operation and scheduling.

Classical mathematical methods and heuristic algorithms are frequently used to solve the optimal scheduling problem of microgrid. The classical mathematical method has advantages in solving speed and convergence [2], but it is easy to fall into local optimization or even fail when dealing with complex nonlinear, discontinuous objective functions and constraints [3,4]. In contrast, the heuristic algorithm is less dependent on the mathematical model and is easier to deal with nonlinear problems, so it has been widely used in different optimization problems of power systems [5], but the parameter setting of the heuristic algorithm is more random and the result is greatly affected by it. Microgrid has flexible composition, a complex operation mechanism, and a large amount of data while operating, however, the above methods do not effectively accumulate and utilize the scheduling knowledge.

Transfer Learning, as an effective means to reuse knowledge, has shown excellent performance in image recognition, text classification, emotion classification, and so forth [6].

However, its application in the field of power systems is still in the exploratory stage. At present, scholars have made achievements in power system supply and demand interactive real-time scheduling [7], power system decentralized carbon energy composite flow optimization [8], economic risk scheduling [9], and so forth. In the above research, TL is frequently combined with Reinforcement Learning (RL) to achieve the purpose of knowledge accumulation and knowledge updating. With deep reused knowledge of TL, RL has been provided strong support. As an important theoretical branch in machine learning, RL has strong abilities of self-learning and memory, in which its agent can interact with the environment to obtain the feedback to guide the action selection, then learn the best strategy and accumulate experience and knowledge. At present, it has been studied in power system security and stability control [10], automatic generation control [11], voltage and reactive-power control optimization [12], optimal power flow control [13], interaction of supply and demand [14], power market [15], power information network [16], and so on. In the microgrid scheduling problem, Liu et al. [17] studied the application of RL in the cooperation of wind power and energy storage. This study shows that RL has good adaptability to the uncertainty and complex constraints of the problem. However, the state and action space are discretized in the study, which leads to errors in the optimization results. Wang et al. [18] and Zhang et al. [19] proposed an economic scheduling model based on RL for the main grid-connected operation and island operation of microgrid, respectively. They used the deep neural network to approximately express the continuous state space, so the error caused by the discretization of the state space and the "Curse of Dimensionality" caused by the excessive state space was improved, but the action space was still discrete, so the best optimal scheduling strategy could not be obtained.

In this paper, we study the microgrid optimal scheduling method based on deep deterministic policy gradient and transfer learning. The optimized scheduling model is proposed, which takes the minimum microgrid operating cost as the objective function. The study includes three parts: (1) the framework and learning process of deep deterministic policy gradient, (2) knowledge transfer rules in transfer learning, and (3) the combination of deep deterministic policy gradient and Transfer Learning. Finally, the feasibility and correctness of methodology was verified in line with simulation, in which Deep Deterministic Policy Gradient (DDPG) extends the traditional RL from discrete action space to the continuous action space. This method can effectively reduce the error caused by discretization of traditional RL, while the actual supply and demand similarity-based TL utilizes the scheduling knowledge effectively.

2. Microgrid System Model and Optimization Model

2.1. Component Model

(1) Solar Power Generation

The solar photovoltaic panel output is given by this expression:

$$P_t^{pv} = \eta^{PV} A_s R_s(t) \tag{1}$$

where P_t^{pv} is the output from solar power generation at time step t; η^{PV} is the conversion efficiency of the solar photovoltaic panel; A_s is the solar photovoltaic panel array area; $R_s(t)$ is the radiation intensity of solar photovoltaic panel at time step t.

(2) Wind Power Generation

The wind power generation output can be approximately expressed by this expression [20]:

$$P_t^{wt} = \begin{cases} P_r \dfrac{V_s - V_{ci}}{V_r - V_{ci}} & V_{ci} \le V_s \le V_r \\ P_r & V_r \le V_s \le V_{co} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where V_s is the wind speed through the wind turbines at time step t; V_{ci} is the start-up wind speed; V_r is the rated wind speed; V_{co} is the cut-out wind speed; P_r is the rated output of wind power generation.

(3) Diesel Generator

As a controllable component, diesel generator can provide electricity when the power supply of uncontrollable components is insufficient, and reduce the dependence of microgrid on the electricity of main grid. The fuel cost model of diesel generator can be approximately expressed by this expression:

$$F(P_t^{die}) = a(P_t^{die})^2 + bP_t^{die} + c \tag{3}$$

where P_t^{die} is the diesel generator output at time step t; a, b and c are the cost factors of diesel generator.

(4) Battery

SOC of battery at each time is determined by the previous moment SOC and exchange power of battery, it can be expressed by this expression:

$$SOC_t = \begin{cases} SOC_{t-1} + \dfrac{\eta P_{t-1}^{ess} \Delta t}{S_{ess}} & P_{t-1}^{ess} > 0 \\ SOC_{t-1} + \dfrac{P_{t-1}^{ess} \Delta t}{\xi \cdot S_{ess}} & P_{t-1}^{ess} < 0 \\ SOC_{t-1} & P_{t-1}^{ess} = 0 \end{cases} \tag{4}$$

where SOC_t is the SOC of battery at time step t; SOC_{t-1} is the SOC of battery at time step $t-1$; P_{t-1}^{ess} is the exchange power at time step $t-1$; $P_{t-1}^{ess} > 0$ and $P_{t-1}^{ess} < 0$ are means battery charge and discharge respectively, $P_{t-1}^{ess} = 0$ denotes battery does not act; η and ξ are the charge and discharge efficiency of battery respectively; Δt is the length of each time step on battery act; S_{ess} is the battery capacity.

In order to ensure the normal operation of the battery and extend its lifetime, the exchange power and SOC are constrained:

(a) Exchange power constraint

$$\begin{cases} 0 < P_{t-1}^{ess} < P_{\text{ch.max}} & P_{t-1}^{ess} > 0 \\ 0 < \left| P_{t-1}^{ess} \right| < P_{\text{dis.max}} & P_{t-1}^{ess} < 0 \end{cases} \tag{5}$$

where $P_{\text{ch.max}}$ and $P_{\text{dis.max}}$ are maximum charge power and discharge power respectively.

(b) SOC constraint

According to the physical limitation on battery, If the battery is over charge or over discharge, it will affect the lifetime of the battery, thus the SOC of the battery needs to be controlled within its own limit. Set $SOC_{\min}$ and $SOC_{\max}$ as the minimum and maximum limited SOC of the battery. The limits of SOC at time step t is given by:

$$SOC_{\min} < SOC_t < SOC_{\max} \tag{6}$$

(5) Load

Load refers to the sum of all kinds of electrical equipment electric power consumed at a certain time, the changing trend of load curve relate to user behavior habits. At time step t, The load can be expressed as P_t^{load}.

2.2. Objective Function

In this paper, the optimization goal is to minimize the microgrid operating cost. The objective function is given by:

$$\min(F_1 + F_2) \tag{7}$$

The F_1 is the fuel cost of diesel generator, F_2 is the transaction cost of the transaction power between the microgrid and main grid.

$$F_1 = \sum_{t}^{T} (a(P_t^{die})^2 + bP_t^{die} + c) \tag{8}$$

$$F_2 = \sum_{t}^{T} (\beta \alpha_t^{buy} \lambda P_t^{grid} \Delta t - (1-\beta)\alpha_t^{sell} \lambda P_t^{grid} \Delta t) \tag{9}$$

where T is the scheduling cycle; α_t^{buy} is the price of purchasing one unit of power from the main grid to microgrid at time step t; α_t^{sell} is the price of selling one unit of power from microgrid to the main grid at time step t; Δt is the scheduling interval; P_t^{grid} is the transaction power between the microgrid and the main grid; $P_t^{grid} < 0$ means microgrid sells power to the main grid, $\beta = 0$; $P_t^{grid} > 0$ means microgrid buys power from the main grid, $\beta = 1$. As Equation (10), P_t^{grid} can be calculated by P_t^{load}, P_t^{pv}, P_t^{wt}, P_t^{die} and P_t^{ess}.

$$P_t^{grid} = P_t^{load} - P_t^{pv} - P_t^{wt} - P_t^{die} + P_t^{ess} \tag{10}$$

The transaction power is calculated by the formula does not include the network loss, which is cannot reflect the actual transaction power, thus this paper considers use the conversion coefficient λ expression the network loss.

3. Optimal Scheduling Method Based on Deep Deterministic Policy Gradient and Transfer Learning

The renewable energy output and load demand are affected by climate and user behavior habits, respectively. Although they have strong uncertainty, the sudden change probability of climate and user behavior habits in the same area or adjacent areas is relatively small. Therefore, the actual supply and demand curve in microgrid on similar days of same area or adjacent areas are very similar. Hence, this paper considers the effective accumulation and utilization of scheduling knowledge through using similarity to provide a priori knowledge for microgrid optimal scheduling. TL can establish knowledge connections for scheduling task groups with similarity; at the same time, the RL strong abilities of memory and self-learning can provide support for the learning, updating, and accumulation of knowledge. When combining with TL, it can realize the effective accumulation and utilization of scheduling knowledge. The method schematic is shown in Figure 1.

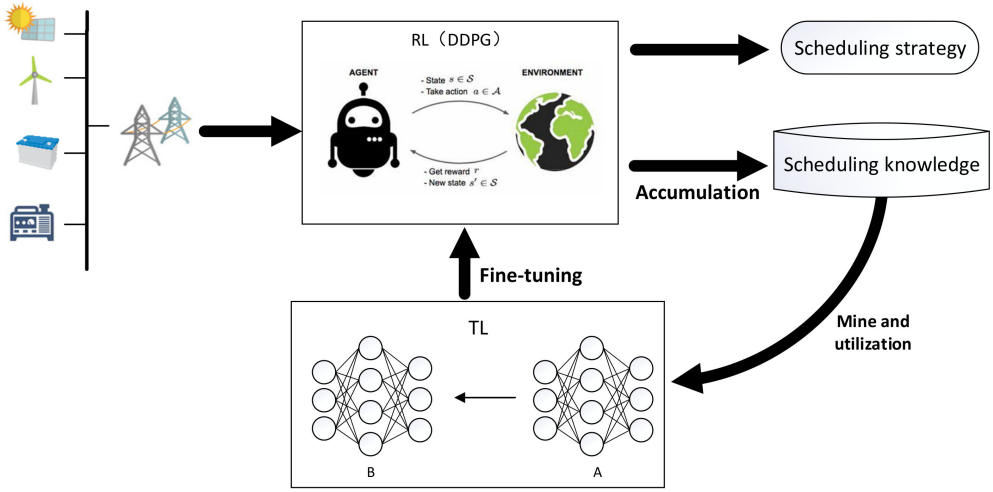


Figure 1. Method schematic.

3.1. DDPG

RL is an artificial intelligence algorithm. In RL, an agent (agent is our artificial intelligence) based on state takes actions within a true or virtual environment, relying on feedback from rewards to find out the foremost suitable policy to achieve its goal. Figure 2 shows the principle of RL.

However, the traditional RL cannot deal with the continuous action space, thus, this paper introduces the DDPG of deep RL as a method to solve the microgrid optimal scheduling problem and combines TL to realize the utilization of scheduling knowledge.

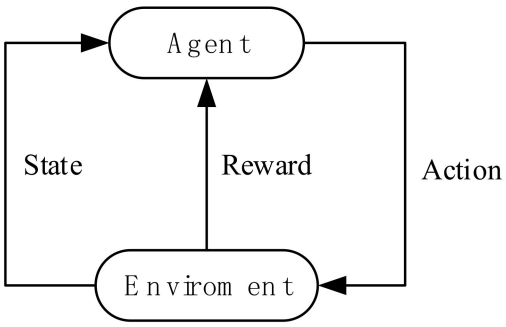


Figure 2. Schematic diagram of reinforcement learning.

DDPG is a policy learning method that integrates a deep learning neural network into Deterministic Policy Gradient (DPG) [21]. DPG is an improved policy learning method based on policy gradient in RL. The policy gradient describes the optimal policy of each step state through the probability distribution function, and the action selection is based on the probability distribution, while the DPG directly obtains the definite value of the decision action at each moment through the policy function, that is,

$$a = \mu(s) \tag{11}$$

The DDPG network structure is shown in the Figure 3, It consists of two parts: the actor network and critic network. DDPG uses the actor network $\mu(s|\theta^A)$ and the critic network $Q(s, a|\theta^C)$ to approximate the policy function $\mu(s)$ and state-action value function $Q(s, a)$ respectively. θ^A and θ^C are the network weights of the actor network and the critic network respectively. The main idea is to generate the action under the guidance of the actor network, and the critic network uses the state-action value function to evaluate the action, then guides the update of its own network and actor network weights through the evaluation.

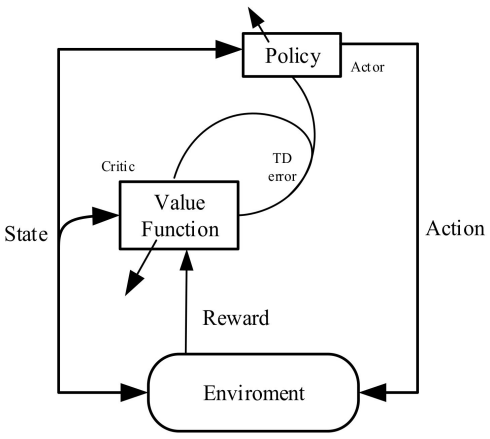


Figure 3. Schematic diagram of Deep Deterministic Policy Gradient (DDPG).

The critic network uses Temporal-Difference to learn the state-action value function, so the loss function of the critic network can be defined as:

$$L(\theta^C) = [Q(s, a|\theta^C) - (r + \gamma(Q(s_-, a_-|\theta^C))]^2 \quad (12)$$

where $Q(s, a|\theta^C)$ is state-action value function obtained by the agent through the critic network, represents the future cumulative reward of the agent after executing the action a in its current state s. As the same, $Q(s_-, a_-|\theta^C)$ represents the future cumulative reward of the agent after executing the action a_- in the next state s_-. All execution actions are generated through the actor network. r is the immediate reward obtained when the agent makes a transition from state s to state s_- perform action a in current time. γ is the discount factor of the cumulative reward value in the future.

The optimization goal of the critic network is given by:

$$\min L(\theta^C) \quad (13)$$

network weights update mode:

$$\theta^C \leftarrow \theta^C + \alpha_C \nabla_{\theta^C} L(\theta^C) \quad (14)$$

The α_C is a scalar step size, called the learning rate of critic network.

The action generated by the actor network is measured by the evaluation of the critic network. The measure function is given by:

$$J(\theta^A) = Q(s, a|\theta^C)|_{a=\mu(s|\theta^A)} \quad (15)$$

The purpose of the actor network is to learn the optimal policy, that the action generated by the actor network can get the maximum cumulative reward value in the future. Therefore, the optimization goal of the actor network is given by:

$$\max J(\theta^A) \quad (16)$$

update weights using the chain rule of gradient:

$$\theta^A \leftarrow \theta^A + \alpha_A \nabla_a Q(s, a|\theta^C)|_{a=\mu(s|\theta^A)} \nabla_{\theta^A} \mu(s|\theta^A) \quad (17)$$

The α_A is a scalar step size, called the learning rate of actor network.

In order to avoid the risk of overestimating, as shown in Figure 4, the DDPG network framework constructed in this paper, adopts the same double network structure as DDQN [22,23], that is, the actor network and critic network simultaneously construct two networks with the same structure but different weights, namely Evaluate net and Target net, The double network structure separates the generation of action a and a_-; the calculation of state-action value $Q(s, a|\theta^C)$ and $r + \gamma(Q(s_-, a_-|\theta^C)$. At the same time, the updating mode of network weights was changed. Evaluate net is updated every time a state transition is performed, and Target net is updated in Soft update [19] mode.

3.2. Knowledge Transfer

3.2.1. TL

TL makes use of the idea of draw inferences about other cases from one instance. TL will effectively use the knowledge learned from the old tasks to similar but different new tasks, so as to improve the utilization of knowledge and the efficiency of new task learning. In TL, the old task is generally called the source domain, and the new task is called the target domain. The knowledge learned in the source domain is affected by the characteristics of the source domain. In the process of knowledge transfer and reuse, the knowledge transfer rules are very important, especially considering the characteristic relationship between the source domain and the target domain. When the knowledge

selection is not appropriate, the knowledge transfer may cause some interference to the target domain, resulting in negative transfer and reduction of learning efficiency in the target domain.

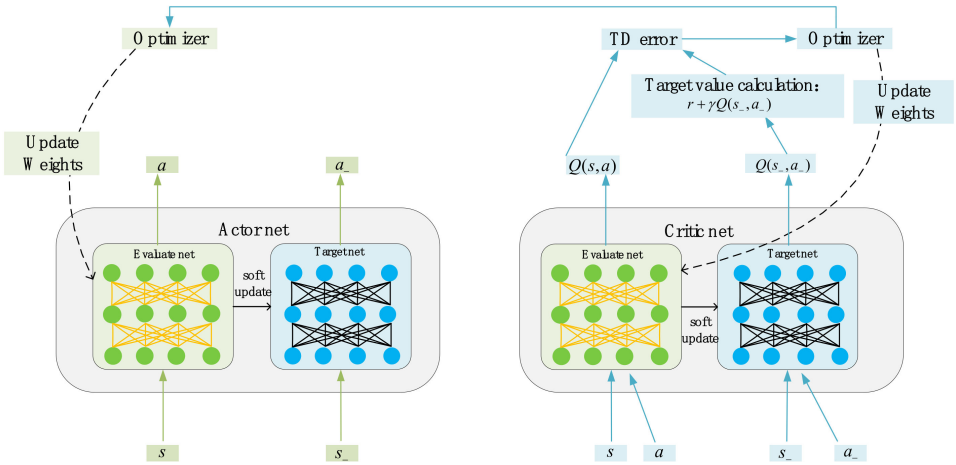


Figure 4. DDPG network structure diagram.

3.2.2. Knowledge Transfer Rules

In the microgrid optimal scheduling, considering the similarity between tasks as the basis for selecting knowledge transfer in the source domain, the rules are formulated as follows:

(1) According to the characteristics of the source domain, an appropriate similarity evaluation function is selected to evaluate the characteristic correlation between the source domain and the target domain.
(2) For the target domain, according to the similarity evaluation function, the similarity between the target domain and the number of N source domains is calculated. The higher the value, the higher the similarity between the target domain and the source domain, which means that source domain knowledge is more instructive to target domain learning.
(3) Selecting the source domain with the highest similarity for knowledge transfer.

3.2.3. Similarity Evaluation Function

On the similarity evaluation function, this paper, we use the inverse number of Euclid Distance as the evaluation similarity function to reflect the actual supply and demand curves similarity between the target domain and source domain. $P^m(t)(m = 1, \cdots, N)$ and $P^{obj}(t)$ denotes the actual supply and demand in N source domains and target domain at each time respectively. The similarity r_m can be calculated by $P^m(t)(m = 1, \cdots, N)$ and $P^{obj}(t)$, as shown in the following Equation (18):

$$r_m = -\sqrt{\sum_{t \in T} \left[P^{obj}(t) - P^m(t)\right]^2} \tag{18}$$

3.3. State-Action Space and Reward Function

3.3.1. State-Action Space

The microgrid optimal scheduling based on DDPG can be formalized as a partially observable Markov decision process, where the microgrid is considered as an agent that interacts with its environment. In this paper, The state space S consists of P^{wt}, P^{pv}, P^{load} and SOC of battery, it can be expressed by:

$$S = \left\{P^{wt}, P^{pv}, P^{load}, SOC\right\} \tag{19}$$

where P^{wt}, P^{pv} and P^{load} are affected by climate and user behavior habits respectively; which are uncontrollable components and can be obtained by prediction. The battery is a controllable component, *SOC* of battery is determined by its own dynamic characteristics, as shown in the Equation (4).

As the controlled components of the microgrid, the operating power of the battery and diesel generator directly affects the scheduling strategy of the microgrid, so the action space is composed of the action power space of the battery and diesel generator. Action space A can be expressed by:

$$A = [[-P_{\text{dis.max}}, P_{\text{ch.max}}], [0, P_{\text{die.max}}]] \tag{20}$$

3.3.2. Reward Function

The effective setting of the reward function can provide correct guidance for the action selection of the agent, in order to obtain the desired goal. The reward function in this paper corresponds to the instantaneous reward at time t, which is obtained by the addition of the operating cost of the microgrid $r1_t(a_t)$ and the penalty $r2_t(a_t)$ caused by the battery violating the constraint.

$$r1_t(a_t) = \begin{cases} -\alpha_t^{buy} \lambda P_t^{grid} \Delta t - a(P_{die}(t))^2 + bP_{die}(t) + c & P_t^{grid} > 0 \\ \alpha_t^{sell} \lambda P_t^{grid} \Delta t - a(P_{die}(t))^2 + bP_{die}(t) + c & P_t^{grid} < 0 \end{cases} \tag{21}$$

$$r2_t(a_t) = \begin{cases} -k \cdot S_{\text{ess}}(SOC_{\min} - SOC_t) & SOC_t \le SOC_{\min} \\ -k \cdot S_{\text{ess}}(SOC_t - SOC_{\max}) & SOC_t \ge SOC_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

The k is the penalty coefficient for violating the constraint.

The instantaneous reward $r_t(a_t)$ is given by:

$$r_t(a_t) = r1_t(a_t) + r2_t(a_t) \tag{23}$$

3.4. Algorithm Flow

The algorithm flow of the microgrid optimal scheduling method proposed in this paper is shown in Figure 5. The whole process consists of two parts: source domain learning and target domain learning, in which source domain learning adopts the DDPG to accumulate microgrid scheduling knowledge, while target domain learning adopts the TL and DDPG to utilize microgrid scheduling knowledge.

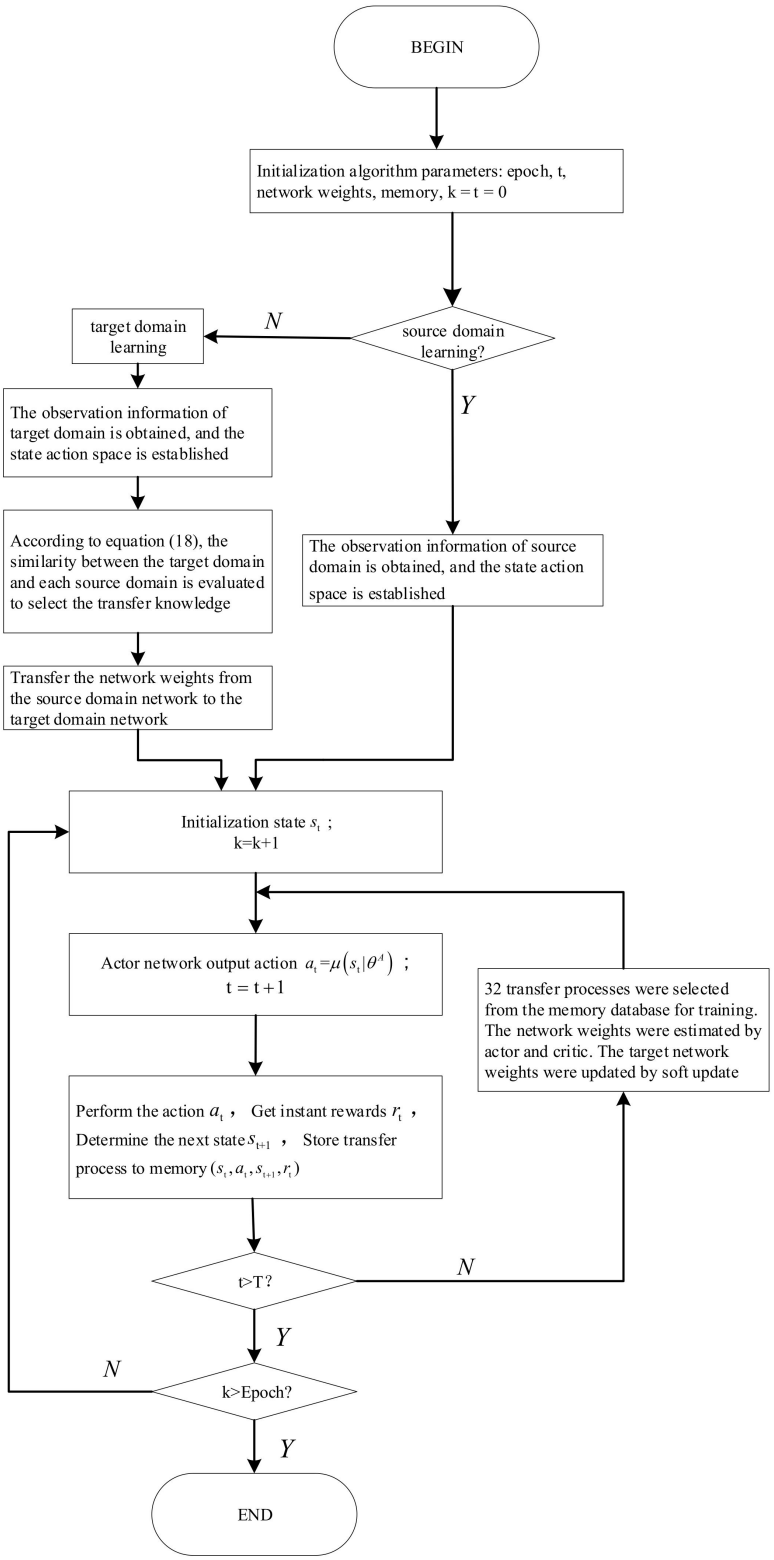


Figure 5. Flow chart of algorithm.

4. Simulation Verification and Analysis

4.1. Simulation

In this paper, solar power generation, wind power generation, diesel generator, battery, load, and energy conversion device are included in the microgrid model, which has an example for simulation. The experimental data of the solar photovoltaic panel output

and load are based on the radiation intensity data and user consumption of GitHub Project [24]. The wind power generation output is based on the wind speed data of Wind Energy Database Project. The capacity of the battery is 175 kWh, the charge and discharge efficiency are 0.9, the maximum exchange power is 30 kW, the minimum SOC of battery is 0.2, the maximum SOC of battery is 0.9, and the initial SOC of battery in this simulation is 0.4. In the DDPG, the actor network has two hidden layers, and they have 50 neurons and 20 neurons, respectively. The activation function is the Rectified Linear Unit (RELU) function. The hidden layer structure of the critic network is the same actor network, in which the variable learning rate and the variable discount coefficient are adopted in the training, and the initial learning rate of the actor network and critic network are set to 0.005. The initial value of discount coefficient factor is 0.9.

The simulation sets up two experiments: source domain learning and target domain learning, which verify the effectiveness of DDPG in continuous action space, and the effective accumulation and utilization of scheduling knowledge based on DDPG and TL, respectively.

Based on the consideration of the actual operation, the electricity price adopts the time-sharing unitary electricity price model [18], is shown as Table 1.

Table 1. Electricity price.

Electricity Buys Price (RMB/kWh)	Electricity Sells Price (RMB/kWh)
1.1	0.85

The neural network input is the microgrid observation information extracted from the experimental data set: the solar photovoltaic panel output, the wind power generation output, load, and the SOC of battery complete the learning of source domain and target domain according to the flow in Section 3.4.

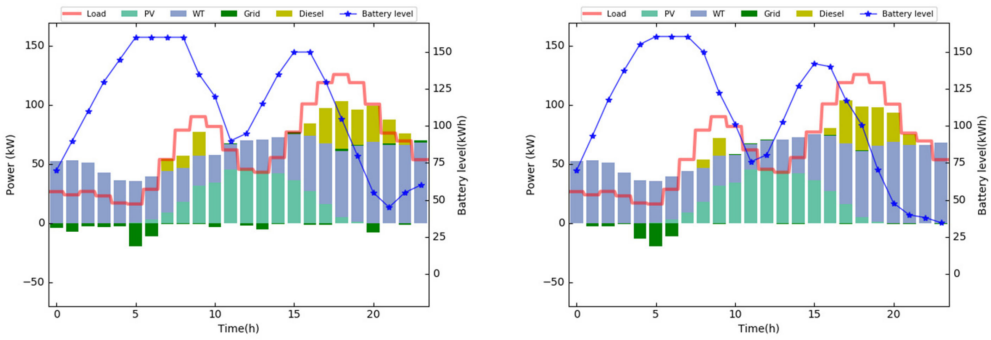
In order to verify the effectiveness on reducing the discretization error, and obtaining excellent scheduling strategy, the proposed method in this paper and the method in [19] are used in the source domain learning experiment. By using the method based on RL in [19], named DDQN, the battery action space and diesel generator action space are discrete, which brings more error because the discrete action space cannot flexibly match the unbalanced power between renewable energy output and load demand. However, by using the proposed method in this paper, named DDPG, both battery action space and diesel generator action space are continuous, which reduces the error because the continuous action space can flexibly match the unbalanced power between renewable energy output and load demand.

(1) DDQN, the power of the battery, and the diesel generator are discretized to 13 and 5 fixed actions respectively, so the action space is set as $A = \{a_1, a_2, \cdots, a_{13\times5}\}$.
(2) DDPG, the action space is set as $A = [[-30, 30], [0, 40]]$, action $a \in A$.

In order to further verify the superiority of transfer learning, we designed a comparative experiment (using TL and without using TL). The best source domain can be obtained according to the knowledge transfer rules in Section 3.2.3. Then, the scheduling knowledge in the best source domain is used for knowledge transfer. In addition, two source domains randomly selected for knowledge transfer are compared to analyze the TL performance in different similarity source domains.

4.2. Source Domain Learning

In the source domain learning, one-year knowledge accumulation is carried out. However, in order to analyze the performance of the scheduling method based on DDPG, this paper takes a typical day as an example to analyze the performance of scheduling strategy based on DDPG. Figure 6 shows the scheduling strategy of a typical day in different methods.



(**a**) DDQN (**b**) DDPG

Figure 6. Typical daily scheduling strategies in different methods.

Figure 7 shows clearly the differences between DDQN and DDPG in battery action, diesel generator action, transaction power. According to the Figures 6 and 7, it can be concluded that during the whole scheduling cycle, the exchange power of the battery and the output of the diesel generator in DDPG are more flexible, and the transaction power between the microgrid and the main grid in DDPG is less than DDQN. Between 0:00–7:00 and 11:00–14:00, the actual supply of renewable energy in the microgrid exceeds the load demand. At this time, neither DDQN nor DDPG have action on the diesel generator, and both DDQN and DDPG have absorbed excess energy by battery charging. When the battery capacity reaches the limit, the battery remains idle in two methods. Compared with the discrete actions in DDQN, the choice of action in DDPG is more flexible, and DDPG also has less trading power than DDQN. Between 7:00–10:00 and 14:00–0:00, the actual supply of renewable energy in the microgrid is lower than the load demand. At this time, both DDQN and DDPG use a battery and diesel generator to meet the energy shortfall. As shown in Figure 7, compared with DDQN, the diesel generator output and the transaction power between the microgrid and the main grid in DDPG are less. This is because the continuous action space improves the flexibility of action selection, enhances the reliability of the microgrid itself, reduces the dependence on the main grid, and further reduces the operation cost of microgrid.

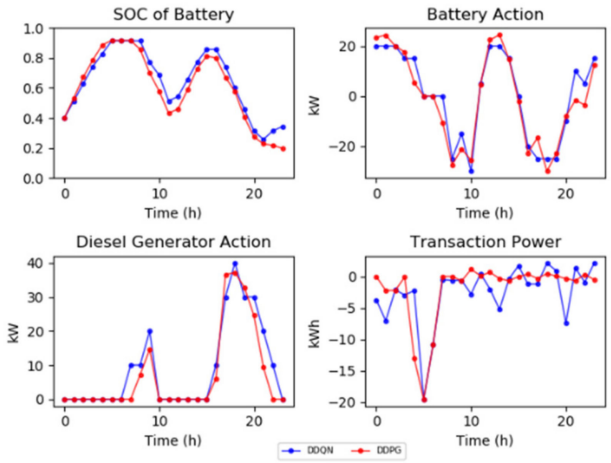


Figure 7. Comparison of typical daily microgrid scheduling strategies in different methods.

It can be seen from Table 2 that in two methods, DDPG obtains the lowest microgrid operating cost: 142.75 RMB. Experiment verifies the effectiveness of the microgrid optimal scheduling method based on the DDPG, and shows that the continuous action space setting can improve the flexibility of action selection, thus reducing the operating cost of the microgrid.

Table 2. Operation income and electricity purchase of microgrid in each method.

Index	DDQN	DDPG
Operating cost of microgrid (RMB)	176.26	142.75
Diesel generator fuel cost (RMB)	118.77	95.50
Transaction cost (RMB)	57.49	47.25
The microgrid buys electricity from the main grid (kWh)	8.92	3.014
The microgrid sells electricity to the main grid (kWh)	68.84	51.06

4.3. Target Domain Learning

In this part, we set the adjacent area scheduling task as the target domain task, and verify the superiority of transfer learning in utilization of scheduling knowledge. The similarity between the target domain and the source domain are evaluated by Equation (18). As shown in Figure 8, the target domain 330 has the highest similarity with the source domain. The source domain 300 is selected for knowledge transfer. At the same time, the other two source domains (source domain 155 and source domain 274) are randomly selected to analyze the performance of TL.

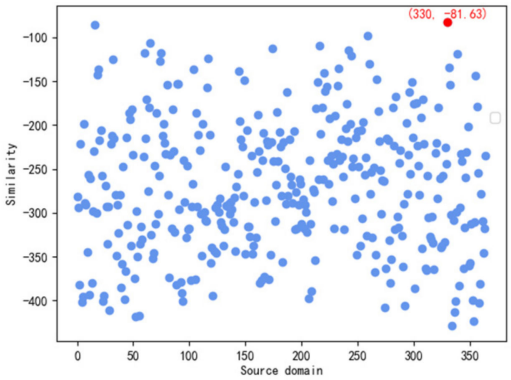


Figure 8. Similarity between target domain and source domain.

Figure 9 shows the scheduling strategy of target domain obtained by target domain learning. During the scheduling cycle, when the output of renewable energy exceeds the load demand, the battery is charged as much as possible within the constraint range; when the output of renewable energy is lower than the load demand, battery discharge cooperates with diesel generator to meet the energy shortfall. In addition, the main grid is also mobilized to absorb the unbalanced power. The scheduling strategy is fully in line with the actual operation, which proves that the microgrid optimal scheduling method based on DDPG and TL proposed in this paper is feasible.

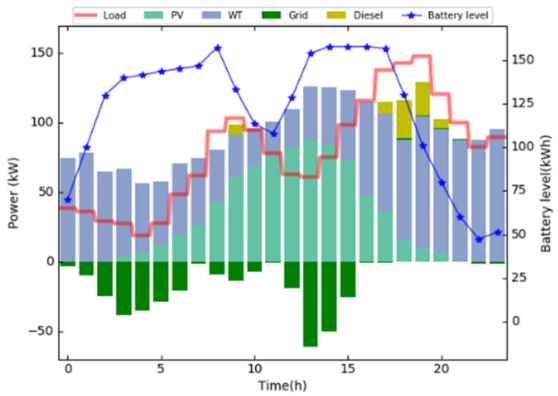


Figure 9. Target domain scheduling strategy.

Figure 10 shows the learning performance of transfer learning for scheduling knowledge in different similarity source domains. It can be observed that when knowledge transfer is not used, learning converges to epoch = 505. When using knowledge transfer, the agent can quickly lock the optimal strategy interval at the initial stage of training. After fine-tuning training, an agent for target domain learning in the source domain 330 with the highest similarity achieves convergence at epoch = 152, while for the agent that carried out the knowledge transfer on the source domain 65 in which the similarity is middle, the relative advantage of convergence rate is small. An agent for knowledge transfer to the source domain 274 with less similarity, the convergence result has deviation; the strategy obtained is inferior to the agent without using TL because the similarity between the target domain and the source domain is low, so the knowledge validity of the source domain cannot be guaranteed. It can be concluded that the similarity between the target domain and the source domain is positively related to the effectiveness of knowledge. The higher the similarity, the higher the effectiveness of knowledge and the better the target domain reuses transfer knowledge.

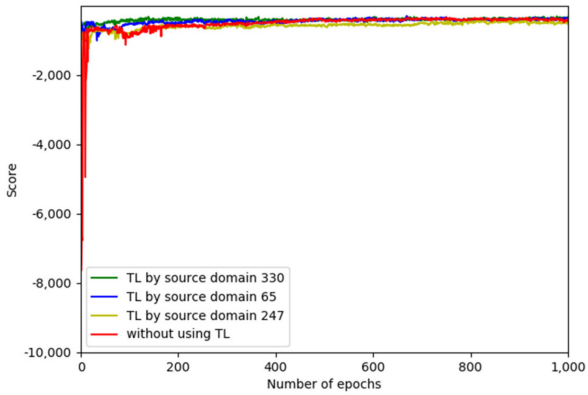


Figure 10. The score curves.

5. Conclusions

Since the optimization methods of microgrid scheduling do not effectively make good use of the scheduling knowledge effectively at present, aiming to solve this problem, this paper proposes a method in which there is optimal scheduling of microgrid based on DDPG and TL.

The findings are listed as follows.

(1) This paper provides an optimal scheduling strategy for microgrid with complex and changeable operation mode flexibility and efficiency through the effective accumulation of scheduling knowledge and the utilization of scheduling knowledge through knowledge transfer.
(2) The DDPG model is introduced into RL, and the action space of traditional RL is extended from discrete space to continuous space.
(3) A microgrid optimal scheduling TL algorithm based on the actual supply and demand similarity is proposed and the effective utilization of scheduling knowledge achieved the transfer of scheduling knowledge.

The scheduling model in this paper does not consider the system power flow constraints and verifies its practicability in large-scale systems; therefore, improving the scheduling model and studying on the state space establishment of large-scale system are the further works.

Author Contributions: Conceptualization, L.F., J.Z., and Y.H.; formal analysis, L.F. and J.Z.; methodology, L.F. and J.Z.; software, L.F., Y.L., and H.Z.; validation, Y.L. and T.H.; writing—original draft, L.F., and T.H.; supervision, J.Z. and Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, Grant Number 51867005. Guizhou Province Science and Technology Innovation Talent Team Project, Grant Number [2018]5615.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

RL	Reinforcement Learning
TL	Transfer Learning
DPG	Deterministic Policy Gradient
DDQN	Double Deep Q Network
DDPG	Deep Deterministic Policy Gradient
RELU	Rectified Linear Unit
diesel	Diesel Generator
t, T	Time indices
pv, PV	Photovoltaic (PV) indices
wt, WT	Wind turbine (WT) indices
Parameters	
N	Number of source domains
η^{PV}	Efficiency of PV
A_{s}	Total area of PV
$P_{die.max}$	Maximum climbing power of diesel generator
P_t^{wt}	Power generated by WT
P_r	Total rated power of WT
V_{ci}	Cut-in speed of WT
V_r	Rated speed of WT
V_{co}	Cut-off speed of WT
S_{ess}	Capacity of battery
η	Efficiency of battery in discharge state
ξ	Efficiency of battery in charge state
$P_{\mathrm{ch.max}}$	Maximum charge power of battery
$P_{\mathrm{dis.max}}$	Maximum discharge power of battery
SOC_{min}	Minimum SOC of battery
SOC_{max}	Maximum SOC of battery
α^{buy}	Electricity buy price of microgrid from main grid
α^{sell}	Electricity sell price of microgrid to main grid
λ	Network loss conversion coefficient
a, b, c	Cost factors of diesel generator
Decision variables	
P^{pv}	Power generated by PV
P^{load}	The load demand
P^{grid}	Transaction power between microgrid and main network
P^{wt}	Power generated by WT
P^{die}	Power generated by diesel
P^{ess}	Change Power of battery
SOC	The state of charge
β	Univariate variable of P^{grid}, $\beta = 1$, $P^{grid} > 0$, $\beta = 0$, $P^{grid} < 0$
P^{obj}	Difference power between renewable energy output and load demand at each time in the target domain
P^m	Difference power between renewable energy output and load demand at each time in m source domain, $m = 1, 2 \cdots, N$
r_m	The similarity between target domain and source domain
V_s	Wind speed
R_s	Solar radiation intensity

References

1. IEEE Power Engineering Society Winter Meeting. In *IEEE Power Engineering Review, Volume PER-4*; IEEE: New York, NY, USA, 1984; p. 18. [CrossRef]
2. Zhang, J.; Fan, L.; Zhang, Y.; Yao, G.; Yu, P.; Xiong, G.; Meng, K.; Chen, X.; Dong, Z. A Probabilistic Assessment Method for Voltage Stability Considering Large Scale Correlated Stochastic Variables. *IEEE Access* **2020**, *8*, 5407–5415. [CrossRef]
3. Dai, C.; Chen, W.; Zhu, Y.; Zhang, X. Seeker Optimization Algorithm for Optimal Reactive Power Dispatch. *IEEE Trans. Power Syst.* **2009**, *24*, 1218–1231.
4. Fumin, Z.; Zijing, Y.; Zhankai, L.; Shengxue, T.; Chenyang, M.; Han, J. Energy Management of Microgrid Cluster Based on Genetic-tabu Search Algorithmy. *High Volt. Eng.* **2018**, *44*, 2323–2330.
5. Ge, S.; Sun, H.; Liu, H.; Zhang, Q. Power Supply Capability Evaluation of Active Distribution Network Considering Reliability and Post-fault Load Response. *Autom. Electr. Power Syst.* **2019**, *43*, 77–84.
6. Zhuang, F.-Z.; Luo, P.; He, Q.; Shi, Z. Survey on Transfer Learning Research. *J. Softw.* **2015**, *26*, 26–39.
7. Xiaoshun, Z.H.A.N.G.; Tao, Y.U. Knowledge Transfer Based Q-learning Algorithm for Optimal Dispatch of Multi-energy System. *Autom. Electr. Power Syst.* **2017**, *41*, 18–25.
8. Zhang, X.; Yu, T.; Yang, B.; Zheng, L.; Huang, L. Approximate ideal multi-objective solution Q(λ) learning for optimal carbon-energy combined-flow in multi-energy power systems. *Energy Convers. Manag.* **2015**, *106*, 543–556. [CrossRef]
9. Xiaoshun, Z.H.A.N.G.; Tao, Y.U. Optimization Algorithm of Reinforcement Learning Based Knowledge Transfer Bacteria Foraging for Risk Dispatch. *Autom. Electr. Power Syst.* **2017**, *41*, 69–77.
10. Tao, Y.; Bin, Z.; Weiguo, Z.H.E.N. Application and development of reinforcement learning theory in power systems. *Power Syst. Prot. Control* **2009**, *37*, 122–128.
11. Ahamed, T.I.; Rao, P.N.; Sastry, P.S. A reinforcement learning approach to automatic generation control. *Electric Power Syst. Res.* **2002**, *63*, 9–26. [CrossRef]
12. Li, T.; Liu, M. Reduced Reinforcement Learning Method Applied to Multi-objective Coordinated Secondary Voltage Control. *Proc. CSEE* **2013**, *33*, 130–139.
13. Sanseverino, E.R.; Di Silvestre, M.L.; Mineo, L.; Favuzza, S.; Nguyen, N.Q.; Tran, Q.T.T. A multi-agent system reinforcement learning based optimal power flow for islanded microgrids. In Proceedings of the 2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC), Florence, Italy, 7–10 June 2016; pp. 1–6.
14. Zhang, X.; Bao, T.; Yu, T.; Yang, B.; Han, C. Deep transfer Q-learning with virtual leader-follower for supply-demand Stackelberg game of smart grid. *Energy* **2017**, *133*, 348–365. [CrossRef]
15. Jiazhi, Z.E.N.G.; Xiongfei, Z.H.A.O.; Jing, L.I. Game among Multiple Entities in Electricity Market with Liberalization of Power Demand Side Market. *Autom. Electr. Power Syst.* **2017**, *41*, 129–136.
16. Li, S.; Wang, X.P.; Wang, Q.D.; Niu, S.W. Research on intrusion detection based on SMDP reinforcement learning in electric power information network. *Electric Power Autom. Equip.* **2006**, *12*, 75–78.
17. Liu, G.; Han, X.; Wang, S.; Yang, M.; Wang, M. Optimal decision-making in the cooperation of wind power and energy storage based on reinforcement learning algorithm. *Power Syst. Technol.* **2016**, *40*, 2729–2736.
18. Yadong, W.; Chenggang, C.; Shensheng, Q. Research on Energy Storage scheduling Strategy of Microgrid based on Deep reinforcement Learning. *Renew. Energy Resour.* **2019**, *37*, 1220–1228.
19. Zhang, Z.; Qiu, C.; Zhang, D.; Xu, S.; He, X. A Coordinated Control Method for Hybrid Energy Storage System in Microgrid Based on Deep Reinforcement learning algorithm. *Power Syst. Technol.* **2019**, *43*, 1914–1921.
20. Zhang, J.; Xiong, G.; Meng, K.; Yu, P.; Yao, G.; Dong, Z. An improved probabilistic load flow simulation method considering correlated stochastic variables. *Int. J. Electr. Power Energy Syst.* **2019**, *111*, 260–268. [CrossRef]
21. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *Comput. Sci.* **2016**, *8*, A187.
22. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing Atari with deep reinforcement learning. In Proceedings of the Workshops at the 26th Neural Information Processing Systems 2013, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 201–220.
23. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
24. François-Lavet, V.; Taralla, D.; Ernst, D.; Fonteneau, R. Deep reinforcement learning solutions for energy microgrids management. In *European Workshop on Reinforcement Learning (EWRL 2016)*; University of Liege: Barcelona, Spain, 2016.