Article

Operation of Distributed Battery Considering Demand Response Using Deep Reinforcement Learning in Grid Edge Control

Wenying Li, Ming Tang *, Xinzheng Zhang, Danhui Gao and Jian Wang

Sichuan Energy Internet Research Institute, Tsinghua University, Chengdu 610042, China; wenyingli2021@126.com (W.L.); Xinzhengzhang2021@126.com (X.Z.); Danhuigao@126.com (D.G.); jianwang20211007@126.com (J.W.)

* Correspondence: Luochuanming2021@outlook.com or tsinghua_energy@163.com

Abstract: Battery energy storage systems (BESSs) are able to facilitate economical operation of the grid through demand response (DR), and are regarded as the most significant DR resource. Among them, distributed BESS integrating home photovoltaics (PV) have developed rapidly, and account for nearly 40% of newly installed capacity. However, the use scenarios and use efficiency of distributed BESS are far from sufficient to be able to utilize the potential loads and overcome uncertainties caused by disorderly operation. In this paper, the low-voltage transformer-powered area (LVTPA) is firstly defined, and then a DR grid edge controller was implemented based on deep reinforcement learning to maximize the total DR benefits and promote three-phase balance in the LVTPA. The proposed DR problem is formulated as a Markov decision process (MDP). In addition, the deep deterministic policy gradient (DDPG) algorithm is applied to train the controller in order to learn the optimal DR strategy. Additionally, a life cycle cost model of the BESS is established and implemented in the DR scheme to measure the income. The numerical results, compared to deep Q learning and model-based methods, demonstrate the effectiveness and validity of the proposed method.

Keywords: grid edge control; demand response (DR); deep reinforcement learning (DRL); multi-agent algorithm; distributed battery energy storage system (BESS); three-phase unbalance

Citation: Li, W.; Tang, M.; Zhang, X.; Gao, D.; Wang, J. Operation of Distributed Battery Considering Demand Response Using Deep Reinforcement Learning in Grid Edge Control. *Energies* **2021**, *14*, 7749. <https://doi.org/10.3390/en14227749>

Academic Editors: Branislav Hredzak and Javier Contreras

Received: 13 October 2021

Accepted: 16 November 2021

Published: 18 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the past ten years, incentive-based DR has developed rapidly in the form of load shedding and transfer, which can greatly improve the flexibility of the grid. According to the National Energy Administration (NEA), BESS capacity will exceed 100 GW·h by 2030 [1]. On the one hand, residential and commercial BESS, etc., are some of the best DR resources, and are increasing sharply with the prosperity of distributed PV, accounting for more than 50% of the newly installed capacity of PV [2]. On the other hand, daily operation of BESS is designed only for the storage of photovoltaic power and saving electricity cost, without consideration of idle use, such as DR programs. Thus, it is necessary to investigate an optimal operation method for BESS in consideration of DR.

However, demand response mechanisms at home and abroad are not yet perfect, and demand response resources are still scarce compared to peak loads. Grid edge control technology provides the potential for exciting transformations in the power industry, creating more choices, higher efficiency, more comprehensive and more efficient decarbonization for customers, and better economic benefits for stakeholders in the value chain [3]. Grid edge control is at the critical point of the adoption curve. Residents, industry and regulators are preparing to connect to digitally distributed resources. Therefore, grid edge control technology could be feasible way to expand resources in a manner that corresponds to demand.

By reducing peak load demand and transferring load demand to low-price and off-peak times, price-based DR is conducive to reducing customer bills [4]. Recent studies

have mainly focused on arranging home power consumption using price-based DR. In [5], Nilsson A et al. controlled smart home appliances using home energy management system (HEMS) and tested the energy-saving potential of smart homes in Sweden. The results showed that the impact of different households on energy consumption varies greatly, indicating that households have a high degree of independence in response to demand. In incentive-based demand response, each customer receives compensation for participating in power demand according to the unified price determined by the utility company [6]. For example, in [7], the authors aimed to maximize the comprehensive benefits of demand response service providers and users, improve the demand response model for power users, and solve the strategy using a deep Q network. In fact, due to the different sensitivities of consumers to prices, a unified price is not able to stimulate all demand response resources. Therefore, ref. [8] established a Stackelberg game model in which a single electricity retailer and multiple users responded based on incentive demand. In this model, when the spot market electricity price was higher than its selling price, the retail e-commerce company developed a demand response subsidy strategy to reduce the loss of electricity sales. In the corresponding period, the user determined the response power according to the subsidy price set by the electricity retailer to obtain additional profits. Incentive-based DR was proposed, acting as an important supplement to price-based DR. However, these papers did not integrate the two DR program. In addition, in the existing literature, DR has always been regarded as an independent procedure, making decisions when there is no information about the operation of the low-voltage distribution network. However, in fact, load reduction or load transfer inevitably change the variables of the station power system, such as active power, increasing its uncertainty. Therefore, it is necessary to integrate the information regarding the power operation of the station area in order to reduce the uncertainty in demand response decision making. For example, in an LVTPA, changes in load cause changes in the three-phase unbalance, which affects the power loss and power quality of the transformer [9].

However, applying conventional algorithms to schedule the appliances of several consumers with mixed-integer variables is difficult due to increased computational complexity and high-dimensional data. For instance, [10] scheduled large-scale residential appliances for DR using a computational algorithm with the cutting-plane method, but did not consider the convergence or calculation time. Recently, a massive number of studies on residential DR have illustrated that reinforcement learning (RL) can tackle the mixed integer programming [11]. The field of deep reinforcement learning (DRL) makes up an active area of research, due to the fact that it is able to find the best implementation strategy by means of “trial and error” in the external environment. Refs. [12,13] implemented batch RL to optimize the operation strategy of heating ventilation and air conditioning, and found an open-loop schedule was required to participate in the day-ahead market. Batch RL can be used to collect empirical data to train the RL agent, and this is popular and critical in areas where it is difficult to build a simulator, or where simulation costs are very high. In addition, the Sara algorithm is an on-policy RL approach, and was applied in thermal comfort control for vehicle cabins based on MDP in [14]. In contrast, Q-learning is an off-policy RL approach that save the learned information in the Q matrix. In [15], the authors focused on the DR of residential and small commercial buildings and studied the optimal energy consumption and storage using Q-learning, the most popular method of RL. Despite of its superior convergence and robustness, Q-learning still suffers from the limitations of continuous states and actions [16,17]. Furthermore, underlying the deep Q-learning networks (DQNs), which combine deep networks and Q-learning by means of the value function [18], the deep deterministic policy gradient (DDPG) algorithm, which is based primarily on the actor–critic method [19], is able to overcome the weakness of DQN in a huge action space. In [20], DDPG was implemented in the state of charge (SOC) management of multiple electrical energy storage systems, and enabled continuous action and smoother SOC control. However, when the application scenario includes a lot of

variables, DDPG performs worse than the multi-agent deep deterministic policy gradient (MADDPG), where each agent represents a demand unit.

In conclusion, the gaps not covered by the current research can be listed as follows.

- (1) The BESS, which is regarded as the best load for DR, is not used fully in the existing DR scheme.
- (2) DR mechanisms and sources are not fully developed, thus limiting the performance of DR programs.
- (3) The current DR models do not consider the relationship between DR and LVTPA.
- (4) There is still a lack of algorithms to solve large-scale DR problems efficiently.

Therefore, inspired by the previous works, this paper proposes a novel DRL-based grid edge controller for BESS, aiming to increase the operation benefits of BESS and three-phase unbalance (TU). The main contributions are summarized as follows:

- (1) Compared to the current DR management, a novel autonomous DR method that considers the BESS within LVTPA is proposed, significantly enhancing the capability of DR.
- (2) The proposed method takes idle utilization of BESS into consideration, rather than normal utilization.
- (3) The proposed method enables the LVTPA to access to information about the local BESS in order to perform global optimization, significantly improving the efficiency of power consumption.
- (4) The proposed DRL-based grid edge controller combines incentive-based edge control with price-based edge control and achieves a better performance compared to conventional methods.

The remainder of this paper is organized as follows. Section 2 details the general structure of the DRL-based grid edge controller and describes the LVTPA. Section 3 formulates the operation problem as a Markov Game, and the optimal strategy is learned by applying the multi-agent deep deterministic policy gradient (MADDPG) algorithm. The numerical study is given in Section 4. Section 5 concludes the paper.

2. Typical Structure of LVPTA for Implementing BESS Demand Response

This section presents the structure of the proposed DRL-based DR controller for residential loads in LVTPA. As shown in Figure 1, a typical LVTPA mainly comprises one low-voltage transformer, one grid edge controller, and several home energy management systems. In a three-phase power system, phase ABC, which carries different loads, especially where the load distribution is unbalanced, easily leads to unbalanced three-phase operation in LVTPA.

Figure 1. The system structure in LVTPA.

2.1. Fuzzification

2.1.1. BESS States

The modeling of BESS is shown in Figure 2 and is mathematically expressed by Expressions (1) and (2) [16].

$$\begin{cases} \text{SOC}(t_{n+1}) = \text{SOC}(t_n) + \frac{\eta_{\text{char}} P_{\text{BESS}}(t_n) \sigma(t_n) \Delta t}{B_r} \\ \sigma(t_n) = 0 \cup 1 \\ \text{SOC}(t_{n+1}) = \text{SOC}(t_n) + \frac{P_{\text{BESS}}(t_n) \sigma(t_n) \Delta t}{B_r \eta} \\ \sigma(t_n) = -1 \end{cases} \quad (1)$$

where P_{rated} is the maximum charge and discharge active power of BESS, P_{BESS} is the current BESS active power, and $S_{\text{soc}, \text{max}}$ and $S_{\text{soc}, \text{min}}$ are the maximum and minimum values of SOC, respectively. $\text{SOC}(t_n)$ is the state of the battery at t_n ; $P_{\text{BESS}}(t_n)$ is the charging and discharging power at t_n . Please note that the charging and discharging operations cannot be performed at the same time. η is the charging and discharging efficiency; B_r is the rated capacity of the battery; $\sigma(t_n)$ represents the state at t_n where 1 represents the charging state, 0 represents disconnection from the power grid, and -1 represents the discharging state.

$$\begin{cases} \sigma(t_{n+1}) = \\ \begin{cases} 0 & \text{SOC}(t_{n+1}) > \lambda_{\text{emg}} \cap h(t_{n+1}) = 0 \\ 1 & \text{SOC}(t_{n+1}) \leq \lambda_{\text{emg}} \cup h(t_{n+1}) = 1 \\ -1 & \text{SOC}(t_{n+1}) > \max(\lambda_{\text{emg}}, \lambda_{\text{dw}}) \cap \\ h(t_{n+1}) = -1 \\ \sigma(t_n) & \sigma(t_n) \neq 0 \cap (t_{n+1} - t_{\text{ini}}) < t_{\text{len}} \end{cases} \end{cases} \quad (2)$$

Figure 2. BESS system model diagram.

In Expression (2), $h(t_n)$ is a decision variable indicating the state of the electric vehicle, where 1 represents the charging state, -1 represents the discharging state, and 0 represents the disconnected state. h is the set of decision variables in which the element is $h(t_n)$. λ_{dw} is the SOC threshold for discharging operation. λ_{emg} is a SOC threshold indicating whether a BESS needs to be charged. If $\text{SOC}(t_n) \leq \lambda_{\text{emg}}$, the charging operation must be carried out immediately. The operational strategy of the battery should ensure that the SOC of the BESS is always greater than this threshold. On the basis of the user's daily night travelling data, the frequency of a user's personal car usage is classified into high-frequency (with λ_{emg} to be 0.9–1.0), intermediate-frequency (with λ_{emg} to be 0.6–0.9), and low-frequency (with λ_{emg} to be 0.25–0.6).

2.1.2. Energy Storage Life Cycle Costs and Benefits

The life cycle costs of energy storage on the user side mainly include the one-time fixed investment cost C_{inv} of the energy storage, and the total operation and maintenance cost C_{ope} ; the benefits include the recovery value B_{rec} at the end of the energy storage life cycle, and the peak and valley arbitrage B_{TOC} of the installation of energy storage during the full life cycle. Demand response total revenue is B_{DR} . F is the full life cycle benefit of energy storage.

$$F = B_{TOC} + B_{DR} + B_{rec} - C_{inv} - C_{ope} \quad (3)$$

Equation (3) is the total operation and maintenance cost of the energy storage.

$$C_{inv} = c_e E_{max} + c_p P_{rated} \quad (4)$$

$$C_{ope} = c_{om} E_{max} \quad (5)$$

In Equations (4) and (5), c_e is the unit capacity cost; c_p is the unit power cost; c_{om} is the annual operation and maintenance cost coefficient per unit capacity; E_{max} is the rated maximum capacity of energy storage; P_{max} is the rated power of energy storage.

The benefits of the entire life cycle of energy storage include energy storage recovery value, full cycle peak and valley arbitrage of energy storage, and demand response benefits, as shown in Equations (6)–(8), respectively.

$$B_{rec} = \theta C_{inv} \quad (6)$$

$$B_{TOC} = \sum_{i=1}^T \sum_{t=1}^{96} c_i (p_{c,i,t} - p_{d,i,t}) \Delta t_i \quad (7)$$

$$B_{DR} = \sum_{l=0}^{l_{DR}} g p_{DR,i} \quad (8)$$

In Equations (6)–(8), θ is the recovery rate of energy storage; c_i is the electricity price at time i ; $p_{c,i,t}$, $p_{d,i,t}$ are the energy storage charge and discharge power of BESS i at time t , respectively. For discharge power, the sampling interval Δt_i is 15 min; T is the total number of days in the full life cycle of the energy storage; $p_{DR,i}$ is the reported demand response power of the energy storage participation of BESS i ; g is the response speed coefficient; l_{DR} is the total number of times that energy storage participates in demand response during the life cycle.

2.2. Model for the Operation of BESS

2.2.1. Energy Storage Operation Constraints

The life loss of energy storage batteries is closely related to throughput, and reducing throughput can prolong service life. In order to make more reasonable use of energy storage, Ref. [21] compared actual user load data with the peak and valley prices of electricity, and found that the daily throughput of energy storage batteries was limited, which not only reduces the throughput of energy storage, it also greatly limits the number of charge–discharge state transitions within a day of energy storage.

$$\begin{aligned} \sum_{i=j+1}^{j+96} p_{c,i} \Delta t_i &\leq m E_{max} (SOC_{max} - SOC_{min}) \\ j &= 0, 1, 2, \dots, n_{day} \\ \sum_{i=j+1}^{j+96} p_{d,i} \Delta t_i &\leq m E_{max} (SOC_{max} - SOC_{min}) \\ j &= 0, 1, 2, \dots, n_{day} \end{aligned} \quad (9)$$

where m is the equivalent number of charge and discharge of energy storage; E_{max} is the rated maximum capacity of energy storage; and SOC_{max} and SOC_{min} are the maximum

and minimum values of the state of charge of energy storage, respectively. The values in this article are 0.9 and 0.1; n_{day} is the number of days.

BESS operates in order to obtain benefits. As shown in Figure 3, these benefits exist in three respects. Firstly, by storing electricity at low-price times and discharging at high-price times, consumers receive the benefit of saving electricity costs. Secondly, by storing electricity at low-price times and discharging to the power grid at high-price times, consumers receive the benefit of spread. Thirdly, by storing electricity when photovoltaic power is sufficient and discharging when there is no photovoltaic power, consumers receive the benefit of saving electricity cost.

Figure 3. Schematic diagram of BESS daily operation.

2.2.2. Physical Constraints of Energy Storage Batteries

$$\text{SOC}_{\min} \leq \text{SOC}_i \leq \text{SOC}_{\max} \quad (10)$$

$$\text{SOC}_{i+1} = \text{SOC}_i + \frac{p_{c,i} \Delta t_i}{\eta^{\text{ch}} E_{\max}} - \frac{p_{d,i} \eta^{\text{dis}} \Delta t_i}{E_{\max}} \quad (11)$$

SOC_i is the state of charge of energy storage at time i ; η^{ch} and η^{dis} are the charging and discharging efficiencies of energy storage, respectively.

$$0 \leq \text{sw}_{c,i} + \text{sw}_{d,i} \leq 1 \quad (12)$$

$$0 \leq p_{c,i} \leq \text{sw}_{c,i} P_{\text{rated}} \quad (13)$$

$$0 \leq p_{d,i} \leq \text{sw}_{d,i} P_{\text{rated}} \quad (14)$$

$\text{sw}_{c,i}$ and $\text{sw}_{d,i}$ are 0–1 variables to indicate the charging and discharging state of energy storage; P_{max} is the rated power of energy storage. Equations (12)–(14) ensure that the energy storage is not in the charging and discharging state at the same time, and that charging and discharging power do not exceed the rated power.

2.2.3. Demand Response Constraints

If energy storage participates in demand response on day t , it must meet the conditions for effective demand response. Equations (15) and (17) restrict the load situation after energy storage has participated in demand response.

$$\max(\text{Load}_k + p_{c,k} - p_{d,k}) \leq \max(\text{Load}_j + p_{c,j} - p_{d,j}) \quad (15)$$

$$\text{mean}(\text{Load}_j + p_{c,j} - p_{d,j}) - \text{mean}(\text{Load}_k + p_{c,k} - p_{d,k}) \geq 0.8 p_{\text{DR}} \quad (16)$$

$$0.05 \text{Load}_{\max}^{\text{pre}} \leq p_{\text{DSM}} \leq 0.2 \text{Load}_{\max}^{\text{pre}} \quad (17)$$

In Formula (15)–Formula (17), k is the response time of the demand response day; j is the corresponding time of the baseline; $p_{c,k}$ and $p_{c,j}$ are the charging power in the corresponding period of energy storage; $p_{d,k}$ and $p_{d,j}$ are the discharge power in the corresponding period of energy storage; Load_k is the load participating in the demand response period; Load_j is the load at the corresponding time 5 days before the response date; p_{DSM} is the optimal response power reported by the user; and $\text{Load}_{\max}^{\text{pre}}$ is the maximum

peak load of the user in the previous year. Equation (15) indicates that the maximum load during the response period does not exceed the baseline maximum load, Equation (17) is the average load constraint during the response period, and Equation (13) restricts the range of the agreed response power.

2.2.4. Power Consumption Constraints of Consumers

In order to satisfy the daily demand of consumers using other stored power, the operation of BESS must consider the following sequence constraints.

$$SOC_{i,t} \geq SOC_{con} \quad (18)$$

$$SOC_{i,t} \geq SOC_{con} \quad (19)$$

where SOC_{con} is the power consumption demand without photovoltaic generation. SOC_{str} is the power storage demand under photovoltaic generation.

3. Methodology

In this section, a Markov Game is designed for DR, and is played by the grid edge controller. The MADDPG algorithm is applied to train the grid edge controller to learn the optimal control strategy. Detailed descriptions are provided as Figure 4.

Figure 4. Grid edge controller schematic diagram.

3.1. Markov Decision Process for DR

A discrete time MDP consisting of three elements, combined with the model presented in this paper, is described below.

3.1.1. State

In this paper, two types of state are defined for the grid edge controller to perform: day-ahead scheduling and real-time load reduction. For day-ahead scheduling, the state is defined as follow:

For price-based DR, the state $s_{i,j}$ is as shown in Formula (20), where T represents the daily time slots, and $p_{i,j}$ is the charge or discharge power at every time slot j of user i . The target state is of the satisfaction of all power consumption demand during one day.

$$s_{i,j}(t) = SOC_{i,j}, j \in T \quad (20)$$

For incentive-based DR, the state $s_{i,j}$ is as shown in Formula (21), where T_{DR} represents the time slots of DR, M is the number of consumers participated in DR, m is the m -th

consumer, and l represents the time slot. The target state is to achieve the load management value during one DR period.

$$s_{m,j}(t_{DR}) = \sum_{l=0}^{t_{DR}} p_{m,j}(l).l \in T_{DR} \quad (21)$$

3.1.2. Action

The grid edge controller also has two types of action, corresponding to the two states as mentioned above. For day-ahead scheduling, the action space is defined as follow:

For price-based DR, the action space is defined as shown in Formulas (8) and (22), where $p_{m,j}$ represents the active power of agent m .

$$a_{i,j}(t) = p_{i,j}(t) \quad (22)$$

For incentive-based DR, the action space is defined as

$$a_{m,j}(t_{DR}) = p_{m,j}(t_{DR}) \quad (23)$$

From the above definition of Formula (20), we can derive:

$$s_{i,j}(t_2) - s_{i,j}(t_1) = p_{i,j}(t_1) = a_{i,j}(t_1) \quad (24)$$

For any positive integer z , when $0 \leq t_0 < t_1 < \dots < t_z$ is satisfied, we have:

$$s_{i,j}(t_1) - s_{i,j}(t_0), s_{i,j}(t_2) - s_{i,j}(t_1), \dots, s_{i,j}(t_z) - s_{i,j}(t_z) \quad (25)$$

Hence, the day-ahead load scheduling is an independent increment process that must satisfy the definition of MDP [22]. In the same way, that the definition of the state and action of incentive-based DR can be proved.

3.1.3. Reward

To guide the grid edge controller to learn the optimal DR strategy, the reward functions for day-ahead scheduling and real-time load reduction are defined as follows.

For price-based DR, reward $r_{i,j}(t)$ is defined as shown in Formula (26), where $\rho(t)$ is the electricity price in time slot y , $t_{m,i}$ represents shiftable appliance working time, and δ is the parameter of TU.

$$r_{s,k}(t) = \rho(t)t_{m,i}a_{i,j}(t)(\delta_{s_{s,k+1},TPU} - \delta_{s_{s,k},TPU}) \quad (26)$$

$\delta_{s,TU}$ is defined as shown in Formula (14), where $I_{A,t}$ is the current of phase A in time slot t . In addition, Formula (27) defines the average current of the three phases.

$$\delta_{s,TU} = (\max(I_{A,t}, I_{B,t}, I_{C,t}) - I_{ave}) / I_{ave} \quad (27)$$

$$I_{ave} = (I_{A,k} + I_{B,k} + I_{C,k}) / 3 \quad (28)$$

For incentive-based DR, reward $r_{r,t}$ is defined as shown in Formula (29), where κ_m is the m -th consumer, and ζ is an adjustable parameter that represents the importance of TU.

$$r_{r,k} = p_{m,j}\kappa_m + (\delta_{s_{s,k+1},TPU} - \delta_{s_{s,k},TU})\zeta \quad (29)$$

3.1.4. Objective Function

This paper presents two objective: day-ahead scheduling and real-time load reduction.

For price-based DR, the objective function of the controller is defined as shown in Formula (30) to maximize the energy benefits and reduce the TU of the LVTPA.

$$\max \sum_k^T \sum_m^M \sum_i^N (\rho_k t_{m,i,k} a_{m,i,k} (\delta_{s_{s,k+1},TU} - \delta_{s_{s,k},TU})) \quad (30)$$

For incentive-based DR, the objective function of the controller is defined as shown in Formula (18) to maximize the DR benefits and reduce the TU of the LVTPA.

$$\max \sum_l^L \sum_m^M \sum_i^N (p_{m,j,l} K_{m,l} + (\delta_{s_{s,l+1},TU} - \delta_{s_{s,l},TPU}) \zeta) \quad (31)$$

3.2. Multi-Agent Deep Deterministic Policy Gradients

In the context of DRL, a computerized agent learns to take actions at a discrete time step t to maximize a numerical reward from the environment, as shown in Figure 5.

Figure 5. Structure of a DRL agent trained with the DDPG algorithm.

In the intelligent DR system designed in this paper with the station area as the core, the energy management system is a single agent that manages the electricity consumption arrangements of every BESS, and communicates with others to obtain information regarding its own actions. The improved features of the MADDPG algorithm are as follows:

- (1) Each agent has its own goals and behavior constraints.
- (2) Each agent can interact with the environment and change the state of the environment.
- (3) Each agent can optimize itself when information is incomplete. The calculation of the whole system is asynchronous and concurrent.

As shown in Figure 6, MADDPG is able to use centralized training and distributed applications to fully improve the optimization efficiency of multiple agents. The detailed pseudocode is as follows in Algorithm 1:

Algorithm 1. MADDPG for grid edge control.

1. Initialize the parameters of critic networks $Q_{i,j}(s, \bar{a}_{i,j} | \Phi_{i,j})$ and the actor networks $\mu_{i,j}(o_{i,j} | \theta_{i,j})$ with weights $\Phi_{i,j}$ and $\theta_{i,j}$ for each agent (i, j) randomly.
2. Randomly initialize each agent target network parameters with critic networks and actor networks $\Phi'_{i,j} \leftarrow \Phi_{i,j}, \theta'_{i,j} \leftarrow \theta_{i,j}$.
3. Clear replay buffer D .
4. **For** episode = 1 to N **do**
5. Randomly initialize a random process χ for every exploration episode.
6. Observe initial state s .
7. **For** $h = 1$ to H **do**
8. Select action $a_{i,j} = \mu_{i,j}(o_{i,j} | \theta_{i,j}) + \chi$ for each agent (i, j) .
9. Execute actions $\bar{a}_{i,j}$ and calculate rewards $\bar{r}_{i,j}$ and next state s' .
10. Store transition $(s, \bar{a}_{i,j}, \bar{r}_{i,j}, s')$ in replay buffer D .
11. Update the state observation $s' \leftarrow s$.
12. **For** agent $(i, j), i=1$ to $I, j=1$ to J **do**
13. Sample a minibatch of K samples randomly. $(s^k, \bar{a}_{i,j}^k, \bar{r}_{i,j}^k, s^{k'})$.
14. Set calculating function $y_{i,j}^k = r_{i,j}^k + \gamma Q_{i,j}^{\mu'}(s^{k'}, \bar{a}_{i,j}^k | \Phi'_{i,j}) \big|_{\bar{a}_{i,j}^k = \mu_{i,j}^k(o_{i,j}^k | \theta_{i,j}^k)}$.
15. Update the critic network by minimizing loss:

$$\zeta_{i,j}(\Phi_{i,j}) = \frac{1}{K} \sum_k \left[Q_{i,j}^{\mu}(s^k, \bar{a}_{i,j}^k | \Phi_{i,j}) - y_{i,j}^k \right]^2$$
16. Update the actor network using the sampled and calculated policy gradient:

$$\nabla_{\theta_{i,j}}(\psi_{i,j}) = \frac{1}{K} \sum_k \nabla_{\theta_{i,j}} \mu_{i,j}(o_{i,j} | \theta_{i,j}) \nabla_{a_{i,j}} Q_{i,j}^{\mu}(s^k, \bar{a}_{i,j}^k | \Phi_{i,j})$$
17. **End for**
18. Softly update the target network parameters for each agent (i, j) :

$$\Phi'_{i,j} \leftarrow \tau \Phi_{i,j} + (1 - \tau) \Phi'_{i,j}$$

$$\theta'_{i,j} \leftarrow \tau \theta_{i,j} + (1 - \tau) \theta'_{i,j}$$
19. **End for**
20. **End for**

Figure 6. Structure of training with the MADDPG algorithm.**4. Case Study****4.1. Simulation Environment**

In this paper, a simulation is carried out on the basis of two-year operation data in a three-phase power system including six households who take part in the DR program. The agents were programmed using a personal computer with six Intel cores i7-8700@3.20GHz, 12 threads, and 16 GB memory. The maximum training episode was set to be 20,000. Python

3.6.2 and Tensorflow 2.1 were adopted to program the MADDPG algorithm. In addition, the adam optimizer was applied to update the parameters of the critic network. Other parameters adopted in the MADDPG algorithm are shown in Table 1. The time-of-use prices are given in Table 2. The incentive-based DR subsidy is 8 CNY/kW·h. In addition, the PV feed-in tariff is 0.6 CNY/kW·h. Table 3 shows the other scenario data and benefits.

Table 1. Main parameter configuration of MADDPG.

Parameter	ε	α_A	γ	α_C	τ	Batch size
Value	0.01	0.0005	0.97	0.001	0.001	1024

Table 2. Time-of-use prices of grid.

Periods	Time	Price (CNY/kW·h)
Peak	08:00–12:00	1.1751
	20:00–23:00	
Flat	12:00–20:00	0.7054
Valley	23:00–08:00	0.3351

Table 3. Scenario data.

Consumer	Phase	Price-Based DR	Incentive-Based DR	Capacity $E_{\max}$ (kW·h)	Charging Power (kW)	Electricity Benefit (CNY)	Optimized Benefit (CNY)	Subsidy (CNY)
I	A	✓	-	4.5	2	−4.32	2.23	-
II	A	✓	✓	7.0	3	7.88	12.21	48
III	A	✓	✓	8.0	4	8.16	15.16	64
IV	B	✓	✓	9.0	3	10.26	16.11	48
V	B	✓	✓	6.0	3	6.89	13.88	48
VI	C	✓	✓	6.5	3	7.18	13.45	48

4.2. Numerical Results Discussion

The following section analyzes the case from three perspectives: benefits, three-phase unbalance, and algorithm performance.

4.2.1. Analysis of Cost

This paper compares three types of BESS in Table 4: lithium ion is the cheapest battery, while also having the shortest cycle charge and discharge times. As a result, its unit energy call cost is the highest. Conversely, lithium iron titanate batteries have the longest service life and the lowest unit energy call cost. However, in this work, the lithium iron phosphate battery was tested in the case study due to its popularity.

Table 4. Cost of BESS.

BESS	Cycle Charge and Discharge Times	Installment Cost (CNY/kW·h)	Maintenance Cost (CNY/kW·h)	Salvage Value (CNY/kW·h)	Charge Efficiency	Unit Energy Call Cost (CNY)
Lithium ion	800	1300	200	300	0.90	1.5
Lithium iron phosphate	3000	2300	200	400	0.93	0.7
Lithium iron titanate	10,000	4000	200	500	0.95	0.37

4.2.2. Analysis of Benefits

Due to space limitations, three of the six consumers are compared in this paper. The scheduling of consumer I is shown in Figure 7. It is noticeable that BESS charges in the valley time at 3 kW. Thus, EV can save CNY 2.46 per hour. BESS charges at a low price during 23:00–03:00 and discharges during the periods 07:00–08:00 and 20:00–23:00. During peak times, BESS is fully charged in the 8:00–12:00 period by the PV. When the electricity price increases during the 20:00–23:00 period, BESS is arranged to supply the household power consumption.

Figure 7. Electricity consumption scheduling of consumer I.

In this LVTPA, the power consumption and the PV power values of the consumers were different. Among them, consumer I did not participate in incentive-based DR. As shown in Figure 8, consumer I had a relatively small photovoltaic capacity compared to its power consumption. Therefore, it has no capacity for incentive-based DR. Without optimization using the MADDPG-based grid controller, consumers I, II, and III only considered storing the remaining electrical energy in BESS and using it in the absence of photovoltaic power generation. With optimization using the proposed method in this paper, BESS was arranged to charge at 00:00–02:00 during the valley price and discharge at 08:00–12:00 and 20:00–23:00 during the peak price. PV still generated power to the grid when photovoltaic power was adequate. Meanwhile, BESS still stored the remaining electricity. Similarly, consumers II and III, as shown in Figures 9 and 10, were charged at low price and discharged at high price, respectively, with 2 kW and 3 kW. The BESS of consumers II and III were able to obtain a subsidy during the incentive-based DR period (12:00–14:00).

Figure 8. Power consumption, PV power, and BESS power of consumer I.

Figure 9. Power consumption, PV power, and BESS power of consumer II.

Figure 10. Power consumption, PV power, and BESS power of consumer III.

As shown in Figure 11, every consumer only acquires a small benefit from electricity cost savings and PV electricity earnings. Using MADDPG in grid edge control, the electricity benefit increases significantly. For example, Figure 11 shows that, for consumer I, the earnings increased from -4.32 to 2.23 . Moreover, participation in incentive-based DR is really valuable for BESS owners, and can bring CNY 48 per day at least.

Figure 11. Daily electricity benefits of consumers.

4.2.3. Analysis of Three-Phase Unbalance

As shown in Figure 12, the MADDPG-based grid controller can adjust the power of BESS to reduce the TU. The degree of TU is decreased by 10.3% and the power loss is reduced by 15.1%, as shown in Figure 12.

Figure 12. TU reduction of the whole LVTPA.

4.2.4. Analysis of Algorithm Performance

As shown in Figure 13, the average reward of each agent keeps increasing throughout the training. After 20,000 rounds of training, the rewards that each agent receives are concentrated around one convergence, which means that the agents have learned to hold the power at an arbitrarily predefined value by adjusting the output of BESS under randomly generated scenarios. This proves that the proposed method has good convergence stability.

Figure 13. The training process of each DDPG agent.

Figure 14 shows the performance comparison of the three main reinforcement learning algorithms, i.e., MADDPG (adopted in this paper), DDPG, and DQN. MADDPG outperforms the other two reinforcement learning algorithms, with faster and more precise convergence.

Figure 14. Convergence of the cumulative mean rewards.

Figures 15 and 16 show a performance comparison with the model-based control method in strong light and low light. The model-based method applies convex optimization techniques to calculate the strategy. Both methods successfully maintain good power scheduling. The performance gap is small, but the average computational time of the proposed method is much shorter than that of the conventional model-based method. Especially in large-scale power systems, the proposed method is more efficient than the conventional model-based method.

Figure 15. Comparison of consumption power between MADDPG and the model-based method in strong light.

Figure 16. Comparison of consumption power between MADDPG and the model-based method in low light.

5. Conclusions

This work proposed a multi-agent grid edge controller based on MADDPG for DR management of distributed BESS in LVTPA. The controller was integrated in an automatic grid edge control system, designed to improve the utilization and increase the DR benefits of consumers. The case study demonstrated that the proposed method is able to reasonably

arrange power consumption, thus increasing the total electricity benefits by 40% and reducing TU by 10.3%. In addition, compared to other RL, the proposed method has a better performance with respect to convergence. Compared to the model-based method, MADDPG possesses the same accuracy and displays a superior calculation time under the two different light conditions. These results demonstrate that the proposed DR program has great potential to motivate residential consumers to participate in the DR program. In future work, the proposed DR program can be extended to multiple appliances, including air conditioners.

Author Contributions: Conceptualization, W.L. and M.T.; methodology, W.L. and M.T.; software, X.Z.; validation, W.L., M.T. and D.G.; formal analysis, W.L. and M.T.; investigation, W.L. and M.T.; resources, W.L. and M.T.; data curation, X.Z.; writing—original draft preparation, M.T.; writing—review and editing, W.L. and M.T.; visualization, X.Z.; supervision, J.W.; project administration, J.W.; funding acquisition, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

Nomenclature

DR	Demand response
DRL	Deep reinforcement learning
MDP	Markov decision process
TU	Three-phases unbalance
BESS	Battery energy storage system
RL	Reinforcement learning
DDPG	Deep deterministic policy gradient
IoT	Internet of things
A	A set of actions of agent
MA	Multi-agent
HEMS	home energy management system
LVS	Low-voltage substation
MPC	Model predictive and control
ET	Environment temperature
i	Serial production-line branch index
j	Day-ahead or real time
h	Hour index
$\psi_{i,j}$	The parameter of policy network
$Q_{i,j}$	Individual Q-network
$Q_{i,j}^{\mu}$	Critic network under policy
$Q_{i,j}^{\mu'}$	Target critic network under policy
SOC	State of charge
WM	Wash machine
ANN	Artificial Neural Network
χ	A Noise Process
O	A set of observations
$R_{i,j}$	Cumulative rewards of an agent
a_i	Action of agent i
γ	Discount factor
$\mu_{i,j}$	Policy of an agent
r	Reward
s	States

p_{rated}	Rated active power
p_s	Active power of standby
R	Real-valued reward function
H	Time slot set $\{1, \dots, t \dots T\}$
D_i	Appliance set $\{1, \dots, a \dots, m\}, i \in B$
MILP	Mixed integer liner programming
DQN	Deep Q-network
D	Experience replay buffer
$\theta_{i,j}$	The parameter of policy network
$\Phi_{i,j}'$	Target parameter of Q-network
$\psi_{i,j}'$	Target parameter of policy network
$\theta_{i,j}'$	Target parameter of policy network
$\zeta_{i,j}$	Loss function of Q-network
$o_{i,j}$	Observation of an agent
ϵ	Probability of ϵ -greedy policy
α_A	Learning rate of policy network
α_C	Learning rate of Q-network
TTU	Transformer terminal unit

References

- Guiding Opinions on Promoting the Development of Energy Storage Technology and Industry, NEA, Beijing. 2017. Available online: http://www.nea.gov.cn/2017-10/11/c_136672019.htm (accessed on 1 June 2021).
- Global Energy Storage Market Outlook for the First Half of 2021, BJX Energy Storage Website, Beijing. 2021. Available online: <https://chuneng.bjx.com.cn/news/20210429/1150160.shtml> (accessed on 1 June 2021).
- Report on Energy Supply and Demand in Canada: 2014 Preliminary, Canada's Ministry Ind., Ottawa, ON, Canada, Tech. Rep. 57-003-X. 2016. Available online: <http://www.statcan.gc.ca/pub/57-003-x/57-003-x2016002-eng.pdf> (accessed on 1 June 2021).
- Chen, Z.; Wu, L.; Fu, Y. Real-time price-based demand response management for residential appliances via stochastic optimization and robust optimization. *IEEE Trans. Smart Grid* **2012**, *3*, 1822–1831. [\[CrossRef\]](#)
- Nilsson, A.; Wester, M.; Lazarevic, D. Smart homes, home energy management systems and real-time feedback: Lessons for influencing household energy consumption from a Swedish field study. *Energy Build.* **2018**, *179*, 15–25. [\[CrossRef\]](#)
- Hu, Q.; Li, F.; Fang, X.; Bai, L. A Framework of Residential Demand Aggregation with Financial Incentives. *IEEE Trans. Smart Grid* **2018**, *9*, 497–505. [\[CrossRef\]](#)
- Guo, K.; Gao, C.; Lin, G.; Lu, S.; Feng, X. Optimization Strategy of Incentive Based Demand Response for Electric Retailer In Spot Market Environment. *Autom. Electr. Power Syst.* **2020**, *44*, 28–35.
- Sun, Y.; Liu, D.; Cui, X.; Li, B.; Huo, M.; Xi, W. Equal Gradient Iterative Learning Incentive Strategy for Accurate Demand Response of Resident Users. *Power Syst. Technol.* **2019**, *43*, 3597–3605.
- Anum, M.; Yang, F.; Dong, J.; Luo, Z.; Yi, L.; Wu, T. Modeling and Load Flow Analysis for Three phase Unbalanced Distribution System. In Proceedings of the 2021 4th International Conference on Energy, Electrical and Power Engineering (CEEPE), Chongqing, China, 22–24 April 2021; pp. 44–48.
- Elghitani, F.; Zhuang, W. Aggregating a Large Number of Residential Appliances for Demand Response Applications. *IEEE Trans. Smart Grid* **2018**, *9*, 5092–5100. [\[CrossRef\]](#)
- Zoltan, V.N. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Appl. Energy* **2018**, *235*, 1072–1089.
- Ruelens, F.; Claessens, B.J.; Vandael, S.; de Schutter, B.; Babuška, R.; Belmans, R. Residential Demand Response of Thermostatically Controlled Loads Using Batch Reinforcement Learning. *IEEE Trans. Smart Grid* **2017**, *8*, 2149–2159. [\[CrossRef\]](#)
- Schmidt, M.; Moreno, M.V.; Schülke, A.; Macek, K.; Mařík, K.; Pastor, A.G. Optimizing legacy building operation: The evolution into data-driven predictive cyber-physical systems. *Energy Build.* **2017**, *148*, 257–336. [\[CrossRef\]](#)
- Brusey, J.; Hintea, D.; Gaura, E.; Beloe, N. Reinforcement learning-based thermal comfort control for vehicle cabins. *Mechatronics* **2018**, *50*, 413–421. [\[CrossRef\]](#)
- Mahapatra, C.; Moharana, A.K.; Leung, V.C.M. Energy Management in Smart Cities Based on Internet of Things: Peak Demand Reduction and Energy Savings. *Sensors* **2017**, *17*, 2812. [\[CrossRef\]](#) [\[PubMed\]](#)
- Shao, M.M.; Liu, Y.B. Reinforcement learning driven self-optimizing operation for distributed electrical storage system. *Power Syst. Technol.* **2020**, *44*, 1696–1705.
- Lu, R.; Hong, S.H.; Yu, M. Demand Response for Home Energy Management Using Reinforcement Learning and Artificial Neural Network. *IEEE Trans. Smart Grid* **2019**, *10*, 6629–6639. [\[CrossRef\]](#)
- Lu, R.; Hong, S.H. Incentive-based demand response for smart grid with reinforcement learning and deep neural network. *Appl. Energy* **2019**, *236*, 937–949. [\[CrossRef\]](#)
- David, S.; Guy, L.; Nicolas, H.; Thomas, D. Deterministic Policy Gradient Algorithms. In Proceedings of the 31st International Conference on Machine Learning, ICML, Beijing, China, 21–26 June 2014.

-
20. Mocanu, E. On-Line Building Energy Optimization Using Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2019**, *10*, 3698–3708. [[CrossRef](#)]
 21. Nayak, C.; Kumar, N.; Ranjan, M. Optimal Design of Battery Energy Storage System for Peak Load Shaving and Time of Use Pricing. In Proceedings of the Second International Conference on Electrical, Computer and Communication Technologies, Coimbatore, Tamil Nadu, India, 22 February 2017.
 22. He, S. *Stochastic Process*; CA: Beijing, China, 2008; pp. 150–218.