energies

MDPI

Article

Agent-Based Energy Sharing Mechanism Using Deep Deterministic Policy Gradient Algorithm

Yi Kuang [1], Xiuli Wang [1,*], Hongyang Zhao [1], Yijun Huang [2], Xianlong Chen [1] and Xifan Wang [1]

[1] School of Electric Engineering Xi'an Jiaotong University, Xi'an 710000, China; kuangyi@stu.xjtu.edu.cn (Y.K.); zhaohongyang2014@stu.xjtu.edu.cn (H.Z.); xlchenee@foxmail.com (X.C.); xfwang@xjtu.edu.cn (X.W.)

[2] State Grid Shanghai Municipal Electric Power Company, Shanghai 200000, China; huangyijun1987@hotmail.com

* Correspondence: xiuliw@xjtu.edu.cn

Received: 4 September 2020; Accepted: 22 September 2020; Published: 24 September 2020

Abstract: Balancing energy generation and consumption is essential for smoothing the power grids. The mismatch between energy supply and demand would not only increase the cost on both sides, but also has a great impact on the stability of the system. This paper proposes a novel energy sharing mechanism (ESM) to facilitate the consumption of local energy. With the help of the ESM, multiple prosumers have an opportunity to share surplus energy with neighboring prosumers. The problem is formulated as a leader–follower framework based on the Stackelberg game theory. To address the aforementioned problems, a deep deterministic policy gradient (DDPG) is applied to solve the Nash equilibrium (NE). The numerical results demonstrate that the proposed method is more stable than the conventional reinforcement learning (RL) algorithm. Moreover, the proposed method can converge to NE and find a relatively good energy sharing (ES) pricing strategy without knowing the specific system information. In short, it is notable that the proposed ESM can be seen as a win–win strategy for both prosumers and the power system.

Keywords: energy sharing; Nash equilibrium; deep reinforcement learning; deep deterministic policy gradient

1. Introduction

The distributed energy resources (DER) have recently substantially increased. End users gradually install self-consumed renewable energy sources' (RES) generation on the consumer side. As a result, a new type of entity has emerged in the grid, namely, prosumers who either act as power consumers or power producers in a certain time period. During a different time interval, the prosumers can be sellers or buyers depending on the electricity pricing and their net power profiles. Therefore, it is feasible to improve the local consumption of DER and the stability of the power system through trading energy among neighboring prosumers. Peer-to-peer (P2P) trading emerged as an energy management mechanism that enables each prosumer to participate in energy trading with other local prosumers [1]. From [2], it is demonstrated that P2P energy trading can facilitate the local balance of DER. In order to alleviate the investment cost in upstream generation, increase network efficiency and energy security, Morstyn T. et al. proposed a P2P energy trading mechanism based on bilateral contract networks [3]. Moreover, prosumers are clustered into virtual microgrids to reduce the total energy cost [4]. In [5], a three-tiered framework including micro-grid balancing, aggregator scheduling and trading optimization is designed to provide a dynamic price signal to assist trading-strategy-making, thereby motivating the efficient utilization of distributed energy resources.

However, due to the volatility of distributed renewable generation and the randomness of the end-user behavior, it poses challenges to the reliable operation of power grid, such as power flow

change, line congestion and voltage flicker. It is supposed to be an effective way to integrate distributed energy resources as a virtual power plant (VPP) [6]. With the advent of computer and communication technology, VPP can promote the local consumption of DER. The improvement in electric power market offers an opportunity for VPP to participate in the wholesale market on behalf of the enrolled participants within its territory [7]. On the other hand, VPP can also interact with its inner prosumers through incentive signals [8]. It is essential to develop an effective and practical mechanism for VPP to hedge against these uncertainties of RES generation and prosumers' behavior. With the development of the sharing economy, a concept of energy sharing has been popular and adopted to balance local DERs. Up to now, great efforts have been made to address the ES issue. Traditionally, a market auction method is applied to deal with the ES problem. However, this method is too complicated and inefficient in practice. As a simpler approach, coordinate control can be regarded as an effective way to address the problem [9]. A controller is necessary under the coordinate control framework. The controller needs to determine a reasonable ES price to facilitate the supply–demand balance. Similarly, a VPP formed through P2P transactions between prosumers is proposed in [10], which is applied to coordinate the local prosumers.

Furthermore, a lot of work has been devoted to the design of the ES operating structure [11–13]. A multiagent-based transactive energy framework can be built to manage the excess supply or residual demand in distributed systems [14]. Specifically, adopting the Stackelberg game to build hierarchical models for decision-making problems in power markets attracts more and more attention [15]. A leader–follower structure is proposed to deal with the pricing problem between VPPOC and prosumers [16]. To solve the NE of the game, most studies assume that the information of lower prosumers is available, which lacks authenticity and comprehensiveness. Actually, due to the heterogeneity of real-word prosumers and difference in their behaviors, it is quite difficult and challenging to build models. Recently, with the rapid evolution reinforcement learning (RL), model-free approaches which do not require prior domain knowledge have achieved great success in decision making processes. To date, the most widely used RL algorithms are mainly Roth-Erev (VRE) learning [17], Q-learning [18–20] and their variants [21–23]. It has been proved that Q-learning algorithm has a better exploration ability than VRE. However, for Q-learning, the estimated value of actions are stored by a table, and the table is updated when interacting with the environment. This is only suitable for problems with a low-dimensional, discrete state and action spaces, and it shows poor convergence ability.

In order to address these issues, Mnih et al. combined the deep neural network (DNN) with Q-learning algorithm to formulate a deep Q network (DQN) model [24], which greatly promoted the development of RL. Recently, research about the application of the DQN algorithm is widespread. The authors in [25] proposed a deep reinforcement learning (DRL)-based EV charging navigation, aiming to minimize the total travel time and the charging cost at the electric vehicle charging station. Besides, some researches have attempted to utilize RL to solve the demand response problems in energy management. In [26], the authors formulate the practical energy management problem as a constrained optimal control problem. Due to the renewable power generation devices and loads having no explicit mathematical model, conventional methods appear to be inapplicable, an agent-based, model-free DRL algorithm is applied to obtain the desired control scheme. In addition, an agent-based energy management system is proposed to reduce the peak load and minimize consumption cost in microgrid [27]. For home energy management, a decentralized method based on multi-agent RL is developed to address the decision-making issue with incomplete environmental information [28]. The broad research about the application of RL in demand response and hybrid energy management systems is reviewed in [18,29,30].

Furthermore, the DQN algorithm has been applied to some complex scenarios such as electric vehicle charge scheduling [31], the optimal management of the operation and maintenance of power system [32], energy trading game among smart grids [33], management of a hybrid energy generation system [34,35], and HVAC optimal control [36]. Although the DQN algorithm is capable of dealing with problems with a high-dimensional state/action space, the space is required to be discrete. On that basis, a novel algorithm based on DDPG is applied to solve the problem with continuous space.

As an alternative, DDPG has received more attention. In [37], a DDPG-based decision-making strategy of adaptive cruising for heavy vehicles is presented, taking stability into consideration. A semi-rule-based decision-making strategy for heavy intelligent vehicles based on the DDPG is proposed in [37]. Moreover, the DDPG algorithm is applied to the joint bidding and pricing problem for a load service entity [38], and to model the strategic bidding of market participant [39].

To date, although traditional RL algorithms are capable of solving the no-cooperative game of incomplete information, they are limited to low-dimensional discrete state/action space and is hard to converge to a relatively good solution. Based on the aforementioned research gap, this paper aims to address the limitation of precious methods and introduce a novel expanding application. We proposed DDPG-based agents that share energy locally under a VPP operation framework based on Stackelberg game theory, aiming to help the VPP operator center (VPPOC) to facilitate the local consumption of DER. The ES pricing problem is formulated as an MDP. Different from the traditional model-based approach, the proposed method in this paper has no need of any model information of prosumers. Finally, detailed case studies are provided to demonstrate the effectiveness of the proposed approach.

The main contributions of this paper are summarized as follows:

- A novel ESM under the VPP operation framework is proposed; the interaction between prosumers and VPPOC is formulated as a non-cooperative game problem based on Stackelberg theory;
- A DRL-based model-free approach is proposed to find the NE for the Stackelberg game without requiring any lower model information;
- The effectiveness and stability of ESM is significantly improved based on the DDPG algorithm. The employment of DNN enhances the performance of the proposed model in the processing problem with high-dimensional continuous space.

The remainder of the paper is organized as follows: the problem formulation is introduced in Section 2. Then, the formulation of ES model is proposed in Section 3. Section 4 presents the process of solving the NE problem based on the DDPG algorithm. Experimental scenarios are implemented in Section 5 to demonstrate the effectiveness of the proposed approach. Finally, conclusions are asserted in Section 6.

2. Problem and Formulation

2.1. System Architecture

In this work, we consider an ESM based on VPP. The framework is shown in Figure 1. The interconnected prosumers can exchange electrical energy through the interconnection infrastructure and a communication network. Each prosumer owns distributed RES generation such as wind generation (WG) and photovoltaic generation (PV), and electrical load. Within the VPP operation territory, each prosumer can be regarded as a buyer and seller of electrical energy. Moreover, the loads of each prosumer are assumed to be elastic, which relates to the electrical price. After obtaining the RES generation and planed load information, the VPP operating center determines the energy sharing price and announces it to prosumers. Then, each consumer optimizes the consuming behavior and aim to maximize the revenue or minimize the load cost according to the price signal. The prosumers can be classified into two categories: one is the supply prosumer that shares the surplus energy, the other needs extra energy to meet the load demand which is called demand prosumer. Actually, there is a deviation between the surplus energy and the load demand. Thus, it is essential to set a reasonable ES price to promote the supply and demand matching.

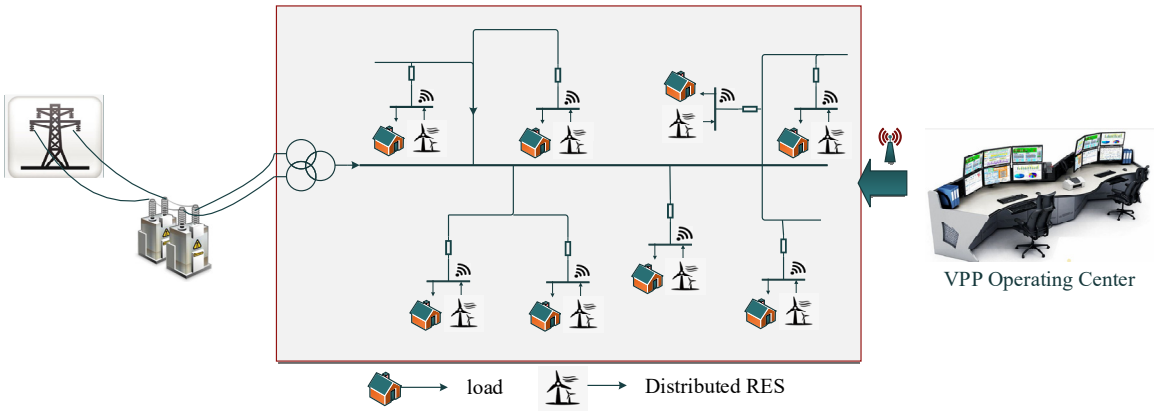

Figure 1. VPP operation framework based on ESM.

2.2. Up-Level: Model

In the hierarchical energy trading framework, VPP is obligated to decide the energy sharing price to reduce the imbalance between surplus energy and load demand. In this paper, the objective of the VPP is to minimize the imbalance energy, which can be donated as

$$\min f^{u}=\sum_{n}^{N}(E_{n,h}^{g}-D_{n,h}) \tag{1}$$

$$\lambda_{h}^{ES,min}\le\lambda_{h}^{ES}\le\lambda_{h}^{ES,max} \tag{2}$$

In (1), f^{u} is the electrical power trade between VPP and electricity market. $E_{n,h}^{g}$ is the power generation of prosumer n at time interval h. $D_{n,h}$ is the actual load demand of prosumer n at time interval h. λ_{h}^{ES} is the energy sharing price which is determined by VPPOC. $\lambda_{h}^{ES,min}$ and $\lambda_{h}^{ES,max}$ are the minimum/maximum energy sharing price limits, respectively. The total energy balance among power market, prosumers and energy storage units can be formulated as follows

$$E_{n,h}^{g}+E_{h}^{m}=D_{n,h} \tag{3}$$

where E_{h}^{m} represents the power energy trading with wholescale power market.

Thus, according to (2), the VPPOC's feasible strategy space can be defined by

$$\Omega_{VPPOC}=\{\lambda_{h}^{ES}\mid\lambda_{t}^{ES}\in\mathbf{R},\lambda_{h}^{ES,min}\le\lambda_{h}^{ES}\le\lambda_{h}^{ES,max}\}\ \ \forall h\in\mathcal{H} \tag{4}$$

2.3. Low-Level Model

In the lower model, when informed of the ES price by the VPPOC, prosumers subscribed to the ES program aim to maximize their incomes or minimize consumption cost. The consumers with surplus RES output attempted to determine their optimal load consumption by taking both the load utility and ES profits into consideration. However, the prosumers demanding extra electric energy try to optimal their load consumption while considering the incurred electrical consumption cost. The objective of the lower problem is donated as

$$\max\ \ f_{n,h}^{l}=\varphi_{n,h}(D_{n,h})+\lambda_{t}^{ES}E_{n,h}^{ES} \tag{5}$$

$$D_{n,h}^{min}\le D_{n,h}\le D_{n,h}^{max} \tag{6}$$

where $f_{n,h}^{l}$ is the incomes of prosumer n at time interval h. $E_{n,h}^{ES}$ represents the energy gap between load consumption and RES output of prosumer n at time interval h. When the value of $E_{n,h}^{ES}$ is positive, it suggests that the prosumer shares surplus electric energy through the ESM. Otherwise, it suggests that prosumers accept extra electric energy from others. $D_{n,h}^{min}$ and $D_{n,h}^{max}$ are the low/up bounds of load consumption. $\varphi_{n,h}(\bullet)$ is the utility function which is used to describe the electric consumption incomes of prosumer n through consuming energy $D_{n,h}$. Specifically, the utility function is used to described the utility of the tasks. In addition, due to the non-decreasing, quadratic

function and logarithmic function are widely used as utility functions [40]. Without loss of generality, this paper adopts the logarithmic utility function, and the detailed formulation of the utility function can be described as follows

$$\varphi_{n,h}(D_{n,h})=\log_{\beta_{n,h}}(1+\alpha_{n,h}D_{n,h}) \quad (7)$$

where $\alpha_{n,h}$ and $\beta_{n,h}$ are parameters varying with prosumer and time. $\beta_{n,h}$ is an experience parameter. Specifically, $\alpha_{n,h}$ is the key factor that can capture the dynamics of prosumer load elastic feature. In this paper, we can set appropriate parameters to demonstrate the impacts of load variation to consumption utility.

According to (6), the feasible consumption strategy space is defined as

$$\Omega_n=\{D_{n,h} \mid D_{n,h} \in \mathbf{R}^{\mathcal{H}}, D_{n,h}^{\min}<D_{n,h}<D_{n,h}^{\max}\} \quad (8)$$

3. Game Theoretic Method

3.1. Stackelberg Game Process

In this section, we reformulate the ES problem as a Stackelberg game. In the aforementioned ESM, the VPPOC is the leader, which sets the ES price by considering the distributed supply–demand energy balance among prosumers, and prosumers are followers which make an optimal load consumption decision according to the price signal and feed it back to VPPOC. The dynamic game process is described in Figure 2.

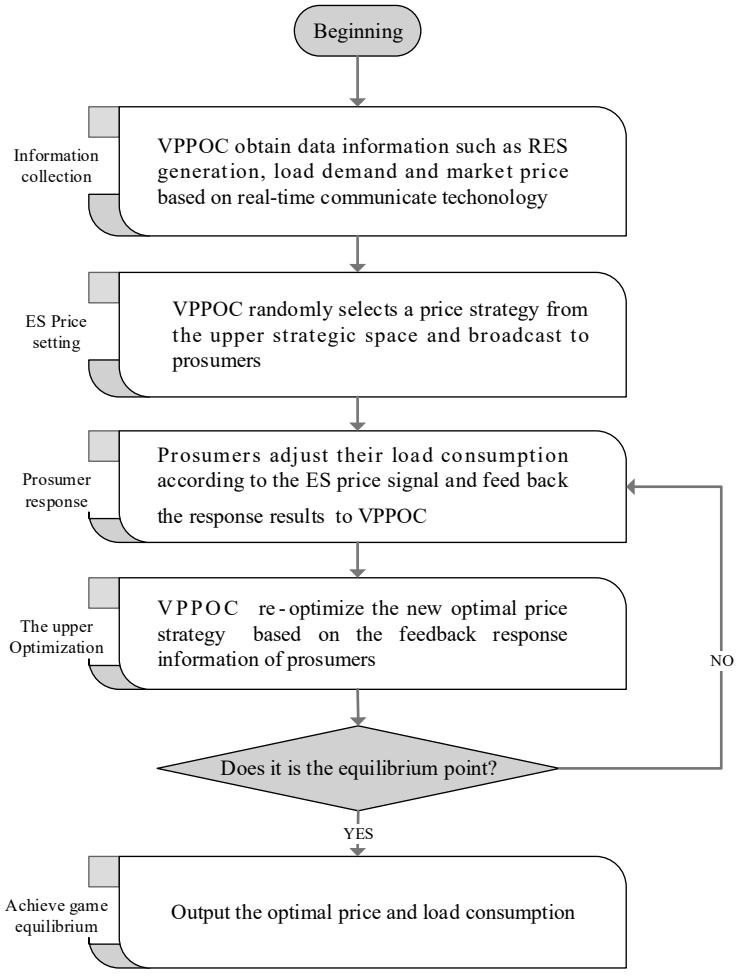

Figure 2. Dynamic game process of Stackelberg.

According to Figure 2, the gaming process is played as the following sequence:

(1) The leader first announces its strategy to followers from the strategy space Ω_{VPPOC}, i.e., an ES incentive price λ_{h};

(2) After informing the pricing strategy λ_{h}, prosumer n chooses a best-response strategy from its strategy space Ω_n as a reaction to the leader VPPOC, which can be viewed as a best-response strategy $\mathcal{D}_{\mathrm{n,h}}(\lambda_{\mathrm{h}})$, which is determined by

$$\mathcal{D}_{\mathrm{n,h}}(\lambda_{\mathrm{h}}) = \arg \max_{\mathrm{D_{n,h}} \in \Omega_n} \mathcal{F}^{\mathrm{l}}_{\mathrm{n,h}}(\mathrm{D_{n,h}}, \lambda_{\mathrm{h}}) \tag{9}$$

(3) Based on the identified best-response strategy set $\mathcal{D}_{\mathrm{h}} = [\mathcal{D}_{1,\mathrm{h}}(\lambda_{\mathrm{h}}), \mathcal{D}_{2,\mathrm{h}}(\lambda_{\mathrm{h}}), ..., \mathcal{D}_{\mathrm{n,h}}(\lambda_{\mathrm{h}})]$ which comprises each of the following prosumers, the leader will select an optimal strategy from Ω_{VPPOC}, donated as $\lambda_{\mathrm{h}}(\mathcal{D}_{\mathrm{h}})$, which can be obtained by solving the upper problem, that is

$$\lambda_{\mathrm{h}}(\mathcal{D}_{\mathrm{h}}) = \arg \min_{\lambda_{\mathrm{h}} \in \Omega_{\mathrm{VPPOC}}} \mathcal{F}^{\mathrm{u}}_{\mathrm{h}}(\mathrm{D_{n,h}}, \lambda_{\mathrm{h}}) \tag{10}$$

(4) At each time interval h, the leader chooses its optimal strategy in (3) and announces it to the follower again. Repeat the above three processes between the VPPOC and the prosumers until the desired NE is obtained.

3.2. Formulation of the Stackelberg Game

The strategic game form is formally defined as

$$\Gamma = \{(\mathcal{N} \cup \mathrm{V}), \{\mathcal{D}\}, \{\mathcal{L}\}, \{\mathbf{F^l}\}, \{\mathbf{F^u}\}\} \tag{11}$$

In dynamic game Γ, a set of follower players $\mathcal{N}$ choose the strategic load consumption strategy from the strategy space $\mathcal{D}_{\mathrm{n,h}}$ to maximize the objective $\mathcal{F}^{l}_{\mathrm{n,h}}$ by considering the ES incentive strategy $\mathcal{L}_{\mathrm{h}}$ determined by the VPP leader V. The leader aims to maximize its objective by optimizing the incentive strategy. Obviously, the game problem is a bi-level optimization problem. We assume the optimal response strategy set as:

$$\mathcal{D}^{*}_{\mathrm{h}} = [\mathcal{D}^{*}_{1,\mathrm{h}}, \mathcal{D}^{*}_{2,\mathrm{h}}, ..., \mathcal{D}^{*}_{\mathrm{n,h}}] \tag{12}$$

Correspondingly, the optimal ES incentive strategy is

$$\lambda^{*}_{\mathrm{h}} \in \mathcal{L} \tag{13}$$

4. Agent Model

In this section, we first formulate the Stackelberg game problem as a finite MDP with a discrete time step. Then, we adopt a DDPG-based model-free approach that does not require full knowledge of the system model information to obtain the equilibrium point of the game problem.

A. Formulation of MDP

At each time interval h, the leader VPPOC agent observes the environment state s_{h}, takes action to determine the ES rate a_{h} and aims to maximize its own cumulative objective. Specifically, the time interval between two adjacent steps is one hour. At time step h, VPP observes the system state, which includes the information about the generation of RES and the load of prosumers. Based on this information, VPPOC chooses an action to adjust the ES behavior. After announcing the ES rates, VPPOC can observe a new environment state and choose a new action; the processes are repeated until an equilibrium point is reached. The detailed MDP decision-making process is illustrated as follows:

(1) State: The system state variable consists of prosumers' load demand and RES generation. The state at time slot h is donated as

$$\mathrm{s}_{\mathrm{h}} = [(\mathrm{D}_{1,\mathrm{h}}, \mathrm{E}^{\mathrm{g}}_{1,\mathrm{h}}), (\mathrm{D}_{1,\mathrm{h}}, \mathrm{E}^{\mathrm{g}}_{1,\mathrm{h}}), \ldots, (\mathrm{D}_{1,\mathrm{h}}, \mathrm{E}^{\mathrm{g}}_{1,\mathrm{h}})] \quad \mathrm{n} \in \mathcal{N}, \mathrm{h} \in \mathcal{H} \tag{14}$$

(2) Action: Receiving the system state s_h at time slot h, VPP operator takes an action a_h, which represents the ES rate λ_h. The action is described as:

$$a_h = \lambda_h \qquad h \in \mathcal{H} \tag{15}$$

The ES rate is constrained by prices bounds as (2), which are derived by a mutual agreement or regulatory requirement between the VPPOC and prosumers, maintaining fair incentive rates and protecting each profit.

(3) State Transition: The states transition from time slot h to time slot h + 1 is shown as follows

$$s_{h+1} = f(s_h, a_h, \omega_h) \tag{16}$$

where the random exogenous factors can influence the state transition. Specifically, the state transition mainly depends on the real-time generation of the RES, which is random and uncertain.

Remark 1. *It is difficult to find an explicit state transition probability from* s_h *to* s_{h+1} *without having a prior knowledge of uncertainties and randomness. In order to set up an accurate distribution model to describe the randomness* ω_h*, a model-free approach based on DDPG is applied to learn the state transition.*

(4) Reward: The reward at time slot h is calculated from the VPPOC perspective as

$$r_h = \sum_{n}^{N} (E_{n,h}^{g} - D_{n,h}) \tag{17}$$

where r_h represents the gap between RES generation and load demand. In an ideal situation, the VPPOC can compensate for the gap as soon as possible via determining a reasonable ES rate (i.e., the total load consumption of prosumers equal to the RES generation).

(5) Action-Value Function: The economic efficiency of VPPOC taking action a under a given state, s, is calculated as the expected discounted cumulative reward from time slot h, which is donated as follows

$$Q^{\mu}(s,a) = \mathbb{E}^{\mu}[\sum_{k=0}^{K} \gamma^k \times r_{h+k} | s_h = s, a_h = a] \tag{18}$$

where $Q^{\mu}(s,a)$ represents action value function; μ is the ES rate policy which maps from the environment state to the ES pricing strategy; γ is the discount factor that balances the importance of the immediate rewards and the future rewards. As the value of γ gets closer to 1, the policy is more foresighted. Conversely, as the value of γ gets close to 0, it indicates that it only takes into consideration the immediate rewards and the policy is shortsighted.

The purpose of the learning process is to find an optimal policy Ω^* over all feasible policies, which maximizes the action value as

$$Q^*(s,a) = \max_{\mu} Q^{\mu}(s,a) \tag{19}$$

where $Q^*(s,a)$ represents the optimal action value function.

The overall diagram of the proposed ESM with DDPG methodology is illustrated in Figure 3. At a separate time interval, VPPOC announces ES price as a leader; the prosumers within its territory determine their load consumption according to the price signal. The interaction between prosumer and VPPOC is fed into the networks. Firstly, the μ network selects an action according to the policy and feeds it into the next Q network. Afterwards, the Q network evaluates the Q value of the action and updates the parameters of the Q function. Finally, an optimal action with a maximal Q value is selected as the ES price and announced to all the prosumers.

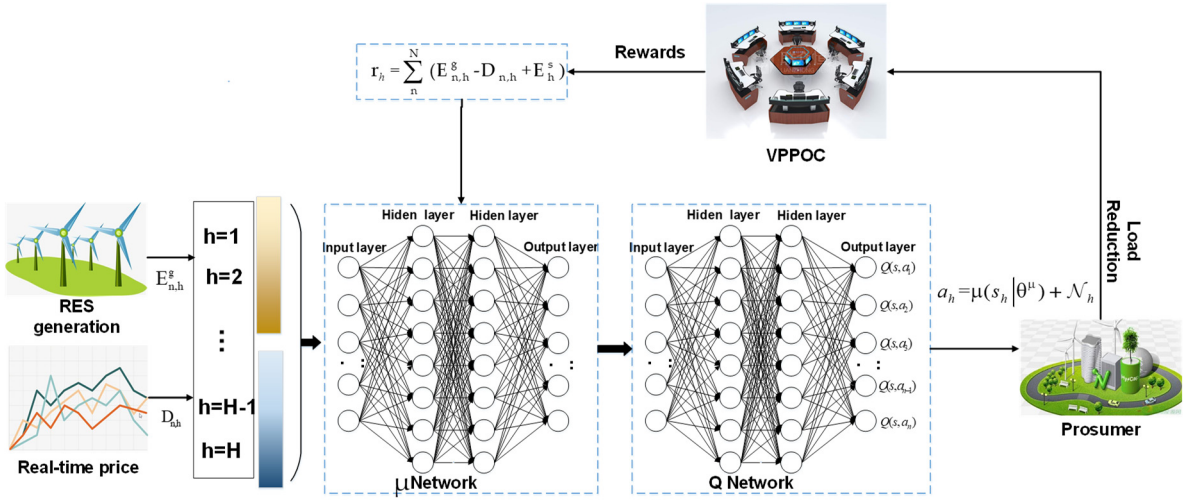

Figure 3. The overall diagram of the proposed decision-making process with deep deterministic policy gradient (DDPG) algorithm.

B. DDPG Algorithm

When the action space dimension is high, especially when it comes to continuous action space, it is difficult for DQN to learn the optimal policy. Thus, DDPG is proposed as an actor–critic algorithm based on the deterministic policy to operate in a problem with a continuous state and action space. DDPG is an Actor–Critic (AC) framework algorithm: it uses a deep neural network as an actor to approximate the policy function, and uses another DNN to approximate the action-value function, which acts as a critic and evaluates the performance of the actor and guides the renewal of the policy network. The actor and critic network architecture is shown in Figure 4.

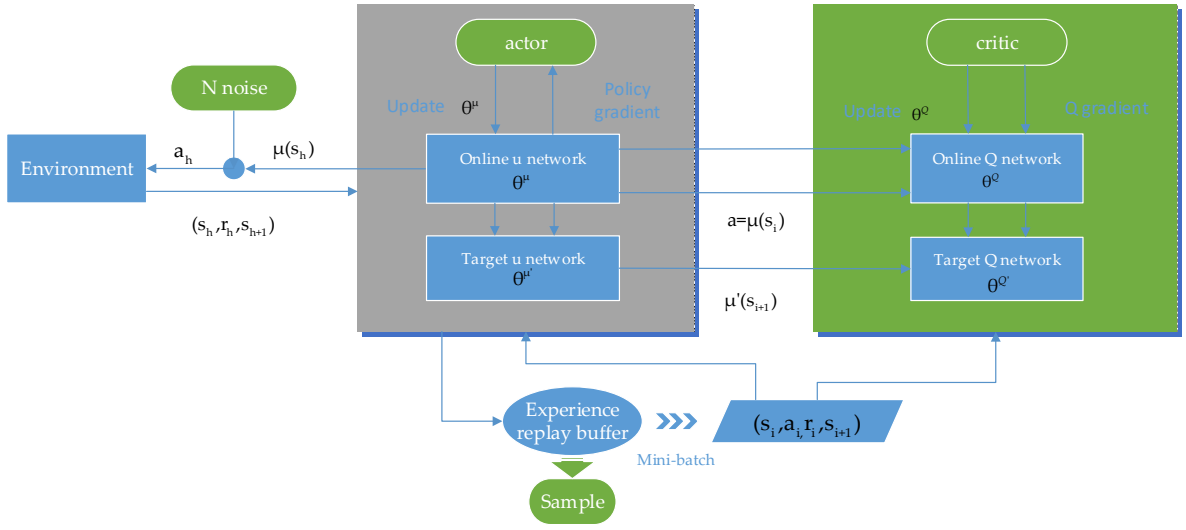

Figure 4. Network structure of DDPG algorithm.

The deterministic behavior strategy of the actor network can be described as a policy function μ with parameter θ^{μ}. The action of each learning step can be obtained through $a_h = \mu(s_h)$. The deterministic behavior strategy of the critic network can be described as an action-value function Q. The parameter of the approximate action-value network is θ^{Q}. Under state s_h, when the agent taking action is a_h according to the actor policy network, the expectation value of reward is known as the Bellman equation [41]:

$$Q^{\mu}(s_h, a_h) = \mathbb{E}^{\mu}[r_h + \gamma Q^{\mu}(s_{h+1}, \mu(s_{h+1}))] \tag{20}$$

The loss of the Q network is shown as follows

$$L(\theta^Q) = \mathbb{E}^{\mu'}[(Q(s_h, a_h \left|\theta^Q\right.) - y_h)^2] \tag{21}$$

where

$$y_h = r_h + \gamma Q(s_{h+1}, \mu(s_{h+1}) \left|\theta^Q\right.) \tag{22}$$

The performance of the policy μ is measured by the performance objective that can be described as

$$J_{\beta}(\mu) = \int_S \varrho^{\beta}(s) Q^{\mu}(s, \mu(s)) ds \tag{23}$$

where the ϱ^{β} is the probability-distributed function of s_h. The target of the training process is to maximize the $J_{\beta}(\mu)$, while minimizing the loss. The actor–critic network training process is described specifically in Algorithm 1.

Algorithm 1 DDPG training process

1: Randomly initialize actor network μ and critic network Q with weights θ^{μ} and θ^{Q}.
2: Initialize target network μ' and Q' with weights $\theta^{\mu'} \leftarrow \theta^{\mu}$, $\theta^{Q'} \leftarrow \theta^{Q}$
3: Initialize experience replay buffer $\mathcal{R}$
4: **for** episode =1:M **do**
5: Initialize a random process $\mathcal{N}$ for action exploration
6: Obtain initial observation state s_1 at time slot *h*
7: **for** h=1:H **do**
8: Select action $a_h = \mu(s_h \left|\theta^{\mu}\right.) + \mathcal{N}_h$ according to the current policy and exploration noise
9: Execute action a_h and observe reward r_h and observe new state s_{h+1}
10: Store transition (s_h, a_h, r_h, s_{h+1}) in experience replay buffer $\mathcal{R}$
11: Sample a random minibatch of N transitions (s_i, a_i, r_i, s_{i+1}) from $\mathcal{R}$
12: Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1} \left|\theta^{\mu'}\right.) \left|\theta^{Q'}\right.)$
13: Update critic by minimizing the loss: $L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i \left|\theta^{Q}\right.))^2$
14: Update the actor policy using the sampled policy gradient:
15: $$\nabla_{\theta^{\mu}} J \approx \frac{1}{N}\sum_i \nabla_a Q(s, a \left|\theta^{Q}\right.)\Big|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^{\mu}} \mu(s \left|\theta^{\mu}\right.)\Big|_{s_i}$$
16: Update the target networks:
17: $$\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'}$$
18: $$\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'}$$
19: **end for**
20: **end for**

5. Results

In this section, multiple case studies are presented to demonstrate the effectiveness of the proposed approach. Firstly, the performance of DDPG in solving the game theoretic problem for a single time interval is demonstrated. Then, we numerically compare the ability of different methods to solve the multi-agent game with incomplete information.

5.1. Experimental Setup

For ease of illustration, simulations were conducted based on six different classify prosumers. The performance of the proposed approach is evaluated based on data from real-work scenarios. The

hourly RES generation and the load demand profiles were obtained on the date of 16 June 2018 from PJM. The training process is carried on the computer with one i7-8700 CPU. The agent-based game theoretic program is trained in Python with Pytorch, a deep reinforcement learning research platform.

5.2. The Evaluation of Training Process

In this part, the DDPG-based approach is trained to solve the multi-agent game problem and find an optimal ES pricing strategy. In order to evaluate the effectiveness of the training process, we first train the agent to learn an optimal ES pricing strategy within a time interval. The evaluation of the cumulative rewards is illustrated in Figure 5. It can be seen that the actions are randomly selected from the action space in the first 2000 iterations. Then, each agent is trained using mini-batch training data that are randomly selected from experience replay buffer $\mathcal{R}$. Finally, the cumulative rewards converge around zero with small oscillations. This result indicates that the optimal action (i.e., ES pricing strategy) promotes the local consumption of prosumers through ES- and DDPG-based approach succeeds in learning a deterministic policy, as shown in Figure 6. The strategic action λ_h converges to λ_h^*. It should be noted that each agent has not been given any information about the other agents, which reflects that the DDPG algorithm can solve the incomplete information game steadily.

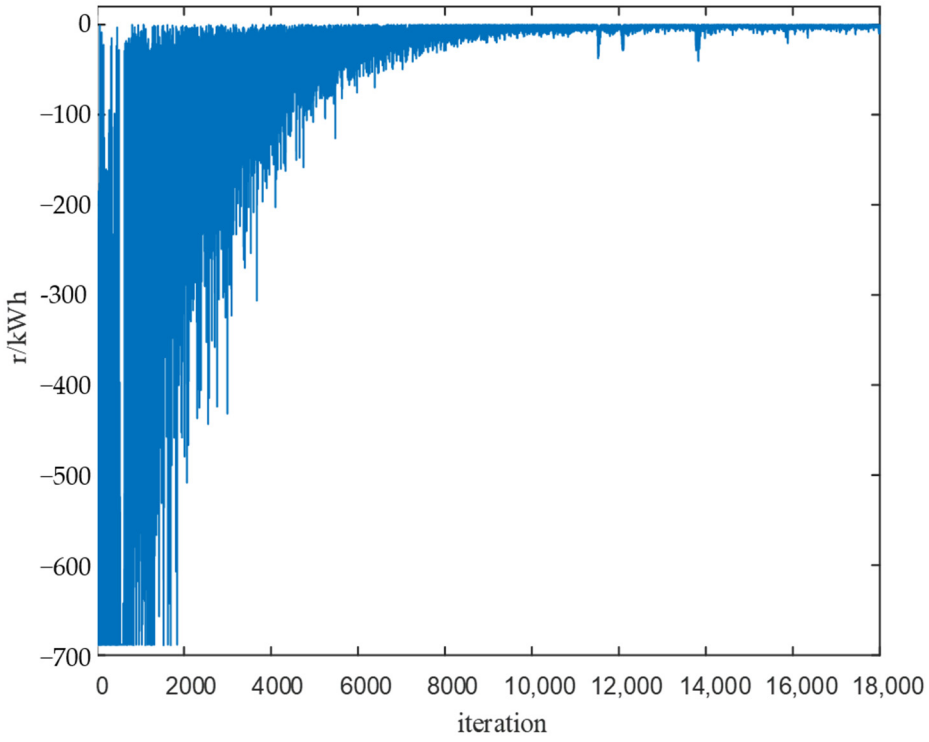

Figure 5. The cumulative rewards during the training process.

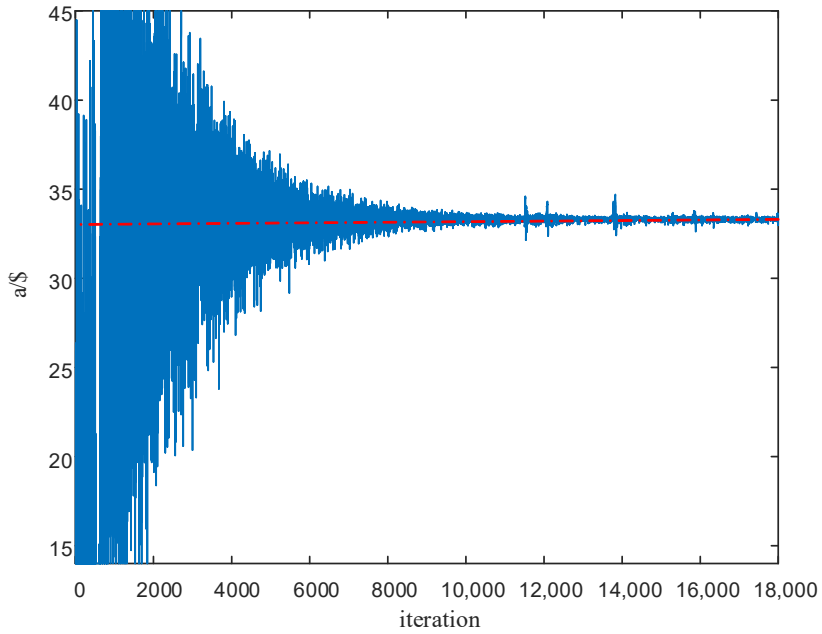

Figure 6. The iterative curve of action.

5.3. The Performance of ESM

Actually, the load elasticity is critical for the stable operation of ES. In order to have a piratical analysis, six different prosumers with different elasticity levels are presented. Elastic coefficient α is proposed to describe the sensitivity of prosumers to the incentive price. Notice that the higher value of α indicates that the prosumers are more sensitive to ES price, i.e., under the same price volatility, the consumption behavior of a prosumer with higher α will change more. It is essential for the stability of ESM to research the relationship between prosumers' consumption behavior and ES price. The ES prices for differently elastic prosumers are illustrated in Figure 7. It is notable that with the increase in elasticity, the ES price that helps to facilitate local energy-demand balance is lower. Hence, having a great understanding of the category of prosumers within the registered area can facilitate VPPOC determining ES pricing strategy reasonably.

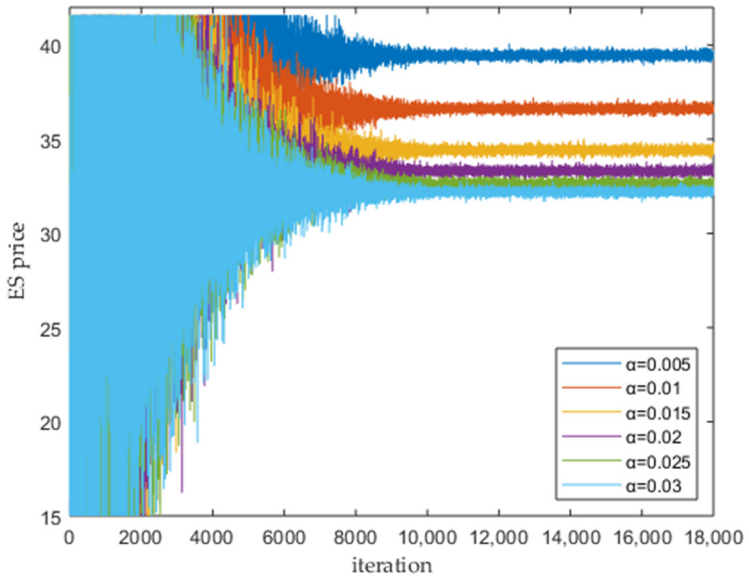

Figure 7. The ES price for different elasticity prosumers.

In fact, for each prosumer, the RES generation and the load demand cannot match exactly. Some prosumers generate more and have surplus energy. However, some consume more and need an extra supply. The detailed ES behaviour of six kinds of prosumers before and after enrolling in ESM is shown in Figure 8. The positive energy represents that prosumers share surplus energy with other prosumers. Conversely, the negative one represents the energy that prosumers received from others. Before enrolling in ESM, only prosumer 3 has surplus energy and the others demand extra supply. The surplus energy cannot satisfy the whole demand. After enrolling in ESM, prosumers reduce their load demand according to the ES price signal. Specifically, the sharing energy of prosumer 2 becomes positive instead of negative. In summary, it can be demonstrated that the total surplus energy can meet the demand as far as possible via the guidance of the ES price signal based on the ESM.

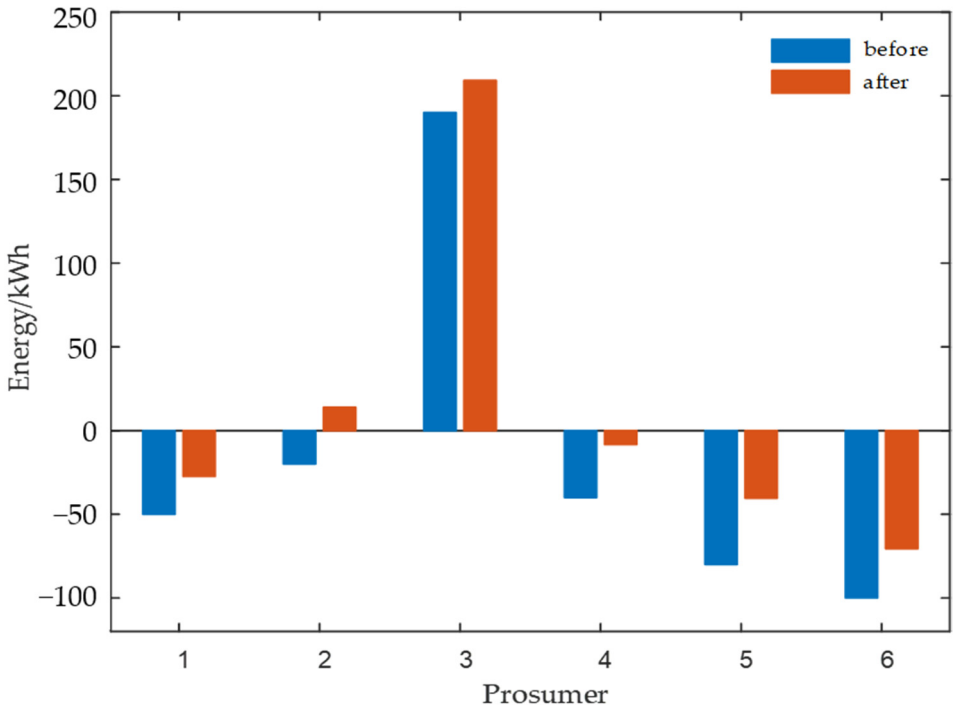

Figure 8. The comparison of renewable energy sources' (RES) generation and load demand before and after enrolling in energy sharing mechanism (ESM).

5.4. Performance of DDPG in Agent-Based Problem

In this section, in order to show the performance of DDPG-based algorithm and verify the ability of an agent to learn and find a reasonable ES pricing strategy, we compare our approach with DQN and analytic iterative algorithm. The action space is continuous in the DDPG algorithm, but is required to be discrete in DQN. Thus, we make different settings for each approach. For DDPG and the analytic method, the action variable λ_h^{ES} is constrained by (2) and can take value arbitrarily within the price bounds. However, action space needs to be reduced in dimension and discretized for DQN. We use discrete nodal price to represent all the prices: $\lambda_h^{ES} \in \{\lambda_{h,1}^{ES}(\lambda_h^{ES,min}), \lambda_{h,2}^{ES}, \ldots, \lambda_{h,m}^{ES}(\lambda_h^{ES,max})\}$. To make the comparison more reliable, the discount factors of DDPG and DQN take the same value.

For ease of illustration, simulation was carried out for 12 typical time intervals that can represent a 24-h day. We assume that there are six category prosumers involved in the ESM. The detailed generation resource type of each prosumer is shown in Table 1.

Table 1. The type of generation resource of each prosumer.

Generation Resource	Prosumer
PV+WG	1,2,5
PV	3
WG	4,6

The original RES and load profiles of six category prosumers are shown in Figures 9 and 10. It is notable that each profile has its own typical feature. Therefore, it can be proved that the proposed approach is practical and enhances the reliability of our conclusion.

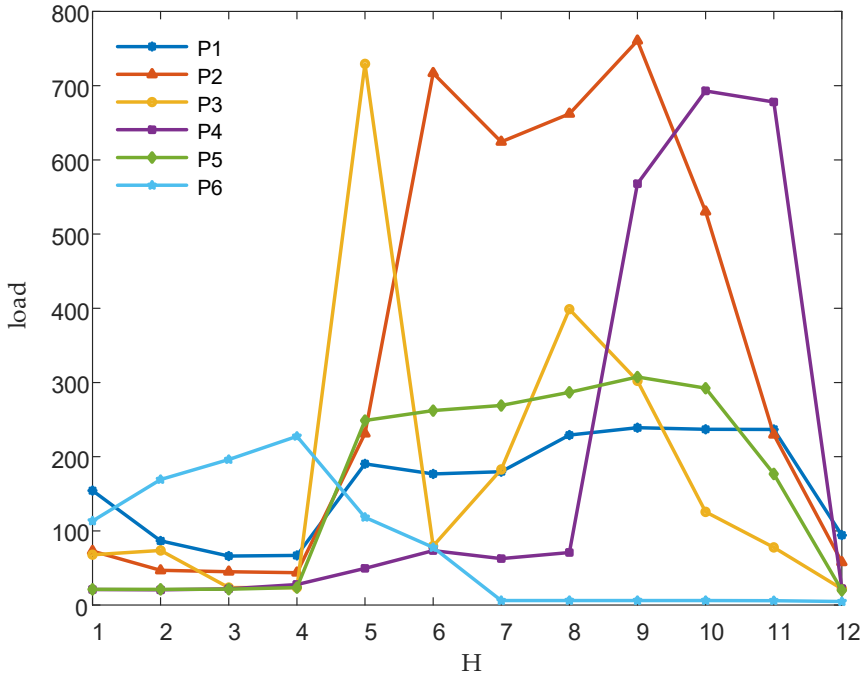

Figure 9. The original load profiles of multiple prosumers.

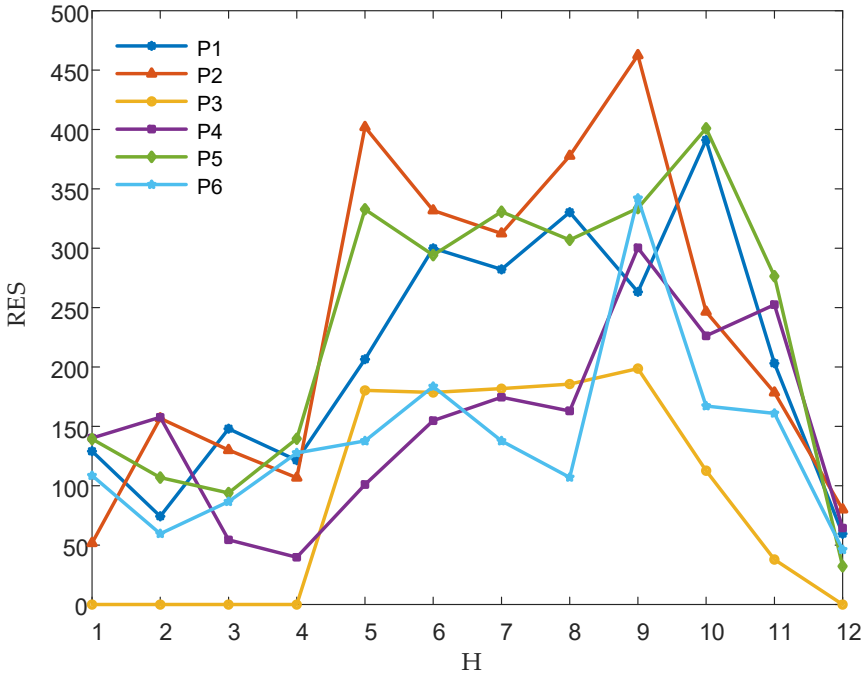

Figure 10. The original RES generation profiles of multiple prosumers.

The converge process of the analytic method, DQN and DDPG is shown in Figure 11. It can be seen that the analytic method obtains the optimal solution successfully. DDPG also converges to optimal solution. These two approaches both work well in solving the game theoretic problem, while DQN does not. Although the learning processes gradually converge, DQN does not find a relatively good solution and have a high volatility for all its iterations. To further compare the ability of three methods, the detailed solving information is listed in Table 2. Note that the analytic method takes the least amount of solving time and iterations to find the optimal solution that is about −11.5kWh. Comparing with DDPG, the total iterations of DQN are only one fifth of DDPG. However, though the solving time of DQN is more than twice as long as DDPG, it still cannot find a relatively good solution. Admitting that the iterations of DDPG are 20,000, it starts to converge after just 3800 iterations and achieves a far smaller error range than DQN. Specifically, as mentioned in Section 2, what makes them superior to the analytic method is that DDPG and DQN can solve the problem without knowing the detailed information of the system and the environment, and DDPG demonstrates a better effectiveness than DQN and is valid for problems with continuous action space.

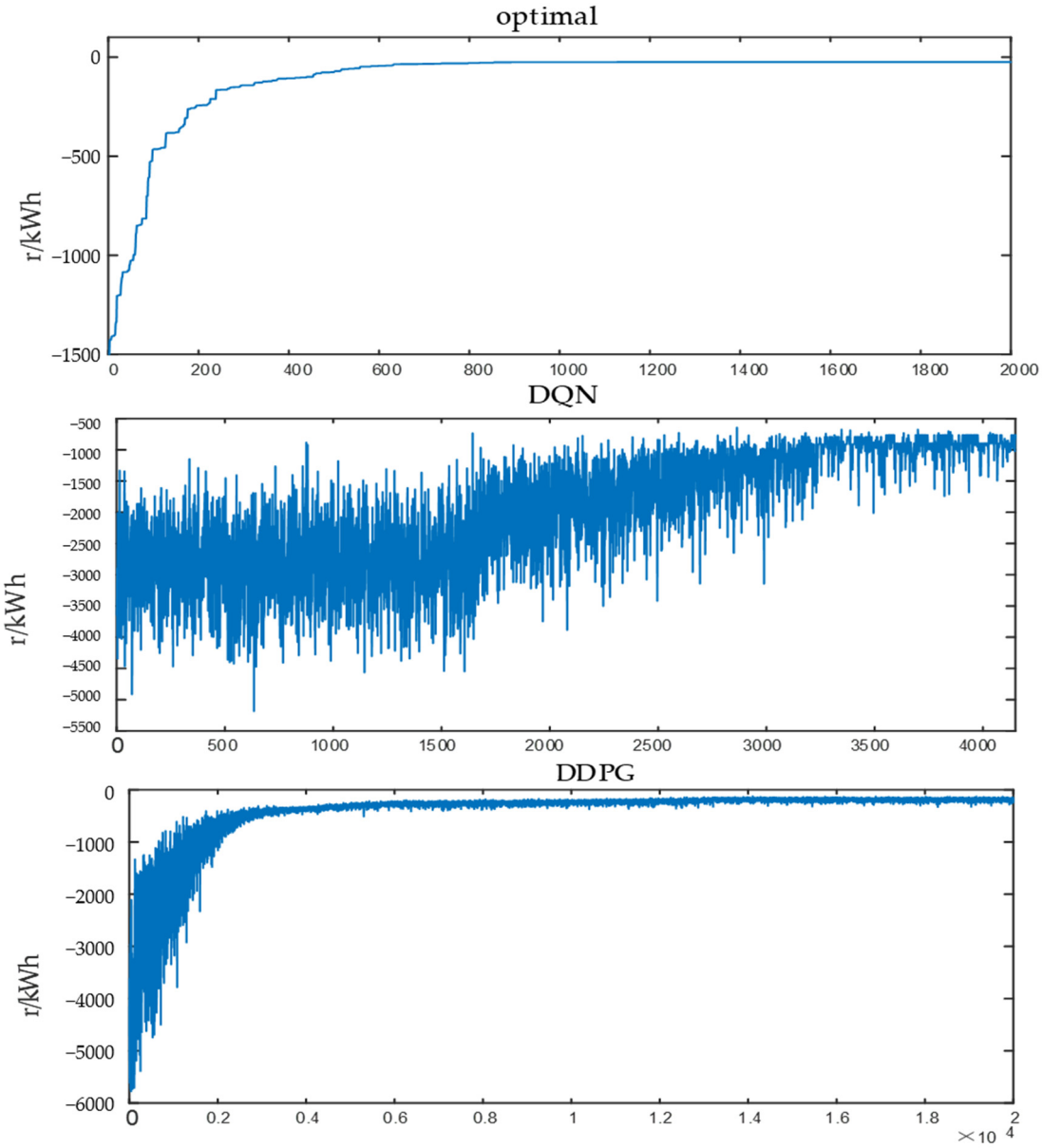

Figure 11. The convergence curve of cumulative rewards under analytic method, deep Q network (DQN), DPG.

Table 2. The solving ability of analytic method, DQN, DDPG.

Method	Total Iteration	Converge Iteration	Solving Time (min)	Result (kWh)
Analytic Method	2000	600	1.25	−11.5
DQN	4000	3300	23	−850~−1450
DDPG	20,000	3800	11	−14.5~−33.6

For further illustration, the ES pricing strategies of three methods are shown in Figure 12. For each time interval, there is a corresponding ES price. It can be seen that the price for each time interval under analytic method and DDPG are quite similar, except for the last time slot. Specifically, the solving error of DDPG is within an acceptable range, while the error of DQN is not. Note that the ES price is relatively low for those time slots at which the total RES generation exceeds the total load demand of all the prosumers. The aim is to incentive prosumers to increase the load consumption. On the contrary, when the RES generation cannot satisfy the load demand, the VPPOC sets a relatively higher price to make prosumers reduce unnecessary load by taking economic efficiency into consideration.

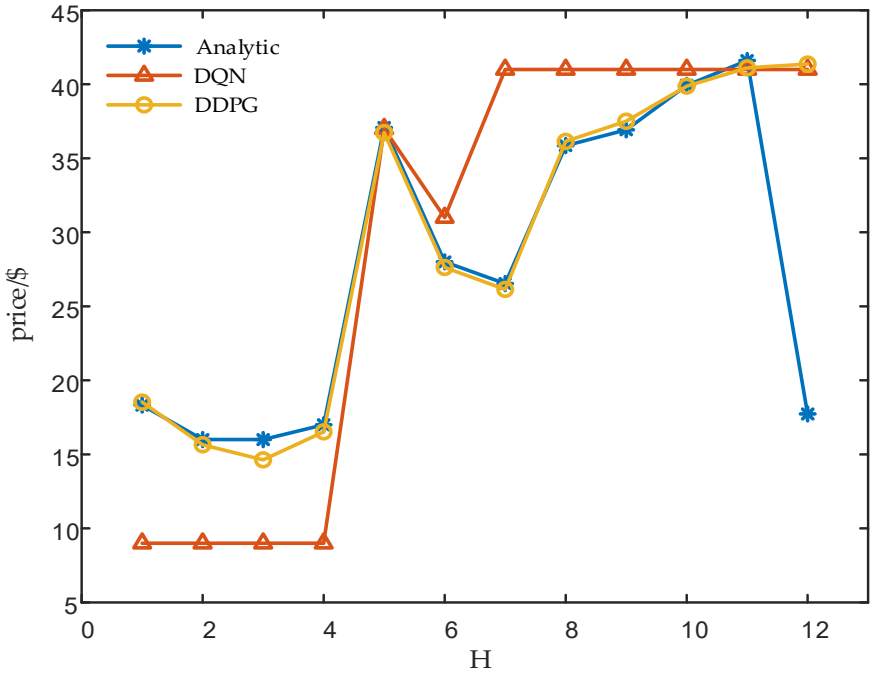

Figure 12. Energy sharing (ES) price at different time intervals under three methods.

5.5. Results of Enrolling in ESM

To verify the effectiveness of our proposed approach, a further case study is presented in the following section. The prosumers have flexibility and the load consumption relates to the price signal. In reality, there is a limit to the elasticity of load that is shown as the dark area in Figure 13. The load variation is constrained within the elastic interval. The load profiles of prosumers before and after participating in the ESM as a VPP are compared in Figure 13. The simulation results show that the proposed ESM can shift the peak load to other time slots. Upon the receipt of the ES price from VPPOC, the prosumers operate the load demand first and feed the result back in real time.

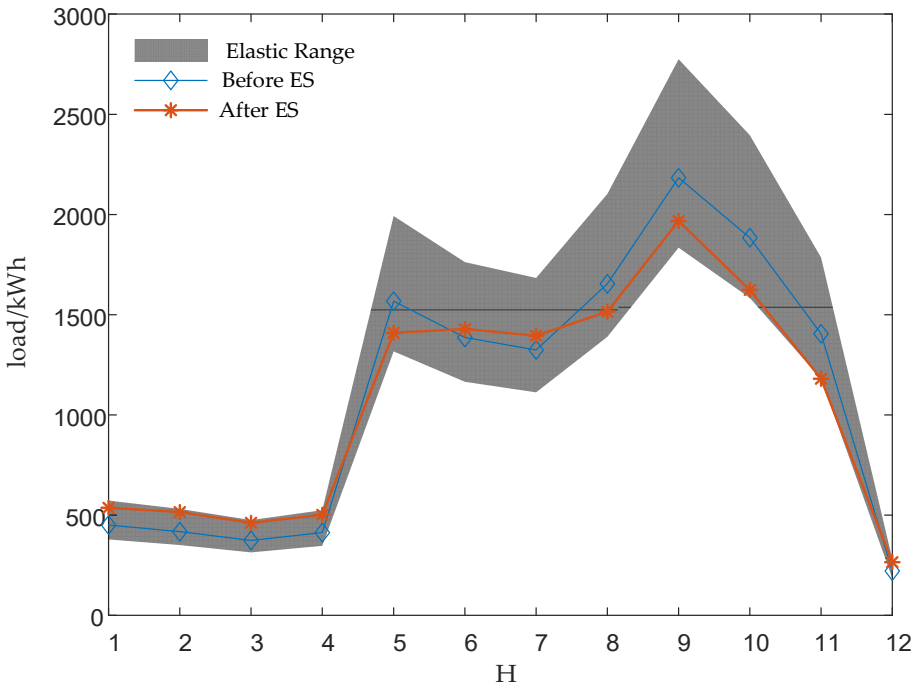

Figure 13. The load elastic range of prosumers, the comparison of load demand profile before and after ES.

Comparing Figures 12 and 13, during the peak load interval (i.e., time slots 8, 9, 10, 11), the RES generation cannot satisfy the prosumers' load demand; due to the higher ES price, prosumers tend to reduce their load demand. Inversely, prosumers will increase power consumption properly during the period of energy surplus; correspondingly, the ES price is low. Therefore, it has more economic benefit and improves energy efficiency by executing the proposed ES mechanism.

The net loads of different prosumers are shown in Figure 14. Positive net loads indicate that prosumers have increased energy consumption and require extra energy through ESM, inversely, the negative net loads indicated that prosumers have surplus energy to share. Note that the diverse load profile, the ESM can be implemented successfully. For the exporting prosumers, the surplus energy can be shared to the neighboring prosumers; they can also receive the sharing energy when they need to. It is of great benefit to utilize energy efficiency. The the gap between supply and demand before and after participating in ESM are compared in Figure 15. Apparently, under the guidance of price signal, it facilitates prosumers' interactions with each other, and supply and demand can be matched locally. Thus, the gap narrows considerably compare to without ES mechanism, which further proves the effectiveness of the proposed method.

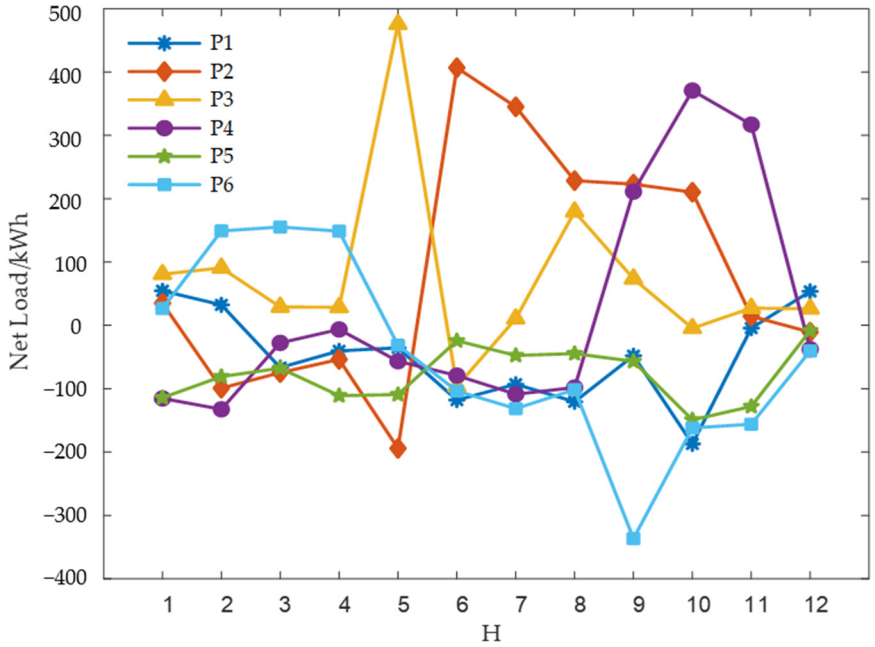

Figure 14. Net load of the prosumers.

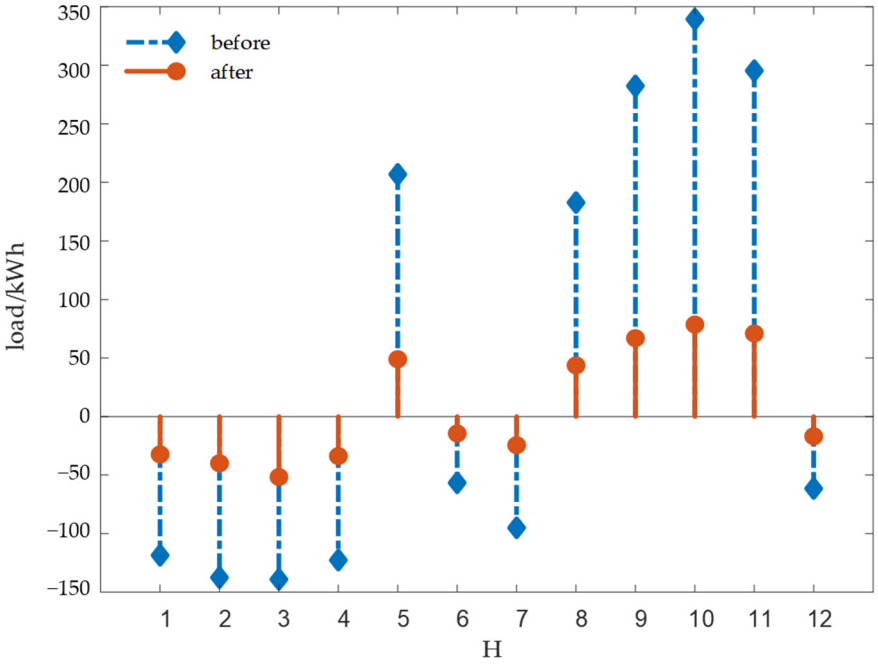

Figure 15. The comparison between supply and demand before and after participating in ESM.

Finally, for each prosumer, the load fluctuation range reduces as shown in Figure 16. It demonstrates that the load profile is smoother after enrolling in ESM and it is important to improve the system reliability and safety, which can be seen a win-win mechanism for both VPP and prosumers.

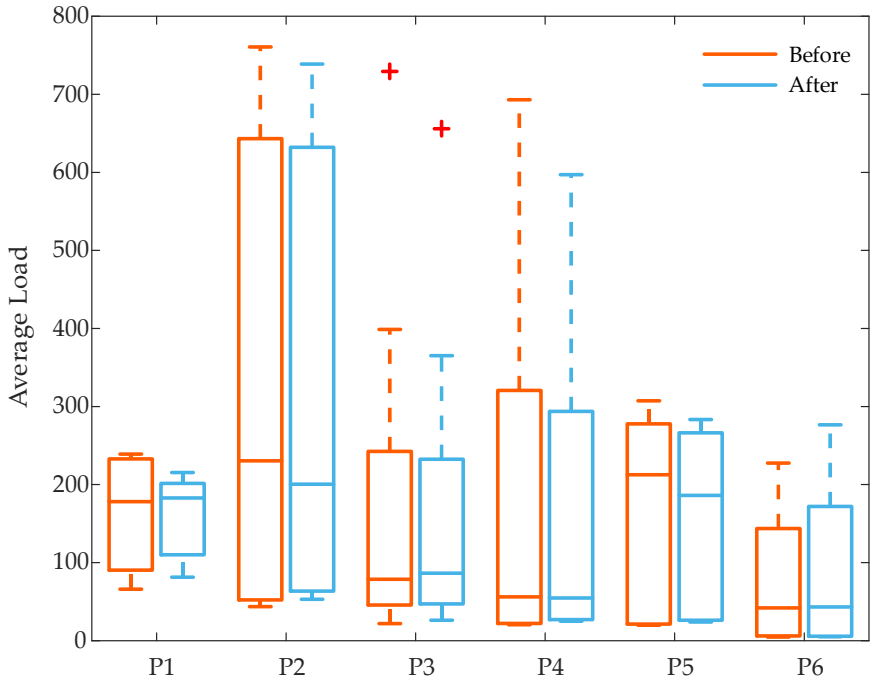

Figure 16. The average load consumption before and after enrolling in ESM.

6. Conclusions

In this paper, we introduce a novel ES simulation model to facilitate the local consumption of DER. VPPOC and prosumers enroll in the ESM under the leader–follower game framework based on Stackelberg theory, wherein the NE of the game can be solved based on the DDPG algorithm. Our model allows VPPOC to derive the optimal ES price without knowing the model information of prosumers; instead, the strategic policy is learned successfully by the agent through dynamic interaction with prosumers. The experimental results illustrate that the proposed ES framework can promote the local consumption of DER, reduce the energy cost for prosumers, balance energy supply and demand within VPP, and improve the stability of the power system.

Due to the proposed agent-based DDPG method showing good convergence in general and having a superior ability to solve problems with a continuous action space, it can be widely used for decision-making problem. Traditionally, the NE needed to solved based on the complete game model information. For our work, the proposed agent-based method is applied to solve the NE under circumstances of incomplete information. It is a supplement to the game theoretic field.

Future work should focus on the following two directions. The first is to apply the proposed ESM in integrated energy system with multiple energies. The second is to apply the DDPG-based agent to demand response management, in a multi-agent game.

Author Contributions: Conceptualization, methodology, writing—original draft preparation, formal analysis, Y.K.; software, H.Z.; validation, X.C. and X.W. (Xifan Wang); project administration, X.W. (Xiuli Wang); funding acquisition, Y.H.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The Design of Electricity Market Mechanism and Key Technology Research under the Responsibility of Renewable Energy Consumption, grant number SGSH0000DJJS2000155.

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. Tushar, W.; Yuen, C.; Mohsenian-Rad, H.; Saha, T.; Poor, H.V.; Wood, K.L. Transforming Energy Networks via Peer-to-Peer Energy Trading: The Potential of Game-Theoretic Approaches. *IEEE Signal. Proc. Mag.* **2018**, *35*, 90–111.
2. Zhang, C.; Wu, J.; Zhou, Y.; Cheng, M.; Long, C. Peer-to-Peer energy trading in a Microgrid. *Appl. Energy* **2018**, *220*, 1–12.
3. Morstyn, T.; Teytelboym, A.; Mcculloch, M.D. Bilateral Contract Networks for Peer-to-Peer Energy Trading. *IEEE Trans. Smart Grid* **2019**, *10*, 2026–2035.
4. Vergados, D.J.; Mamounakis, I.; Makris, P.; Varvarigos, E. Prosumer clustering into virtual microgrids for cost reduction in renewable energy trading markets. *Sustain. Energy Grids Netw.* **2016**, *7*, 90–103.
5. Liu, Y.; Zuo, K.; Liu, X.A.; Liu, J.; Kennedy, J.M. Dynamic pricing for decentralized energy trading in micro-grids. *Appl. Energy* **2018**, *228*, 689–699.
6. Zamani, A.G.; Zakariazadeh, A.; Jadid, S. Day-ahead resource scheduling of a renewable energy based virtual power plant. *Appl. Energy* **2016**, *169*, 324–340.
7. Hooshmand, R.; Nosratabadi, S.M.; Gholipour, E. Event-based scheduling of industrial technical virtual power plant considering wind and market prices stochastic behaviors—A case study in Iran. *J. Clean. Prod.* **2018**, *172*, 1748–1764.
8. Kulmukhanova, A.; Al-Awami, A.T.; El-Amin, I.M.; Shamma, J.S. Mechanism Design for Virtual Power Plant with Independent Distributed Generators. *IFAC-PapersOnLine* **2019**, *52*, 419–424.
9. Luo, Y.; Itaya, S.; Nakamura, S.; Davis, P. Autonomous cooperative energy trading between prosumers for microgrid systems. In Proceedings of the 39th Annual IEEE Conference on Local Computer Networks Workshops, Sydney, Australia, 8–11 September 2014; pp. 693–696.
10. Morstyn, T.; Farrell, N.; Darby, S.J.; McCulloch, M.D. Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants. *Nat. Energy* **2018**, *3*, 94–101.
11. Wang, J.; Zhong, H.; Wu, C.; Du, E.; Xia, Q.; Kang, C. Incentivizing distributed energy resource aggregation in energy and capacity markets: An energy sharing scheme and mechanism design. *Appl. Energy* **2019**, *252*, 113471.
12. Kusakana, K. Optimal Peer-to-Peer energy sharing between prosumers using hydrokinetic, diesel generator and pumped hydro storage. *J. Energy Storage* **2019**, *26*, 101048.
13. Asimakopoulou, G.E.; Dimeas, A.L.; Hatziargyriou, N.D. Leader-Follower Strategies for Energy Management of Multi-Microgrids. *IEEE Trans. Smart Grid* **2013**, *4*, 1909–1916.
14. Nunna, H.S.V.S.; Srinivasan, D. Multiagent-based transactive energy framework for distribution systems with smart microgrids. *IEEE Trans. Ind. Inform.* **2017**, *13*, 2241–2250.
15. Lu, Q.; Lü, S.; Leng, Y. A Nash-Stackelberg game approach in regional energy market considering users' integrated demand response. *Energy* **2019**, *175*, 456–470.
16. Liu, N.; Cheng, M.; Yu, X.; Zhong, J.; Lei, J. Energy-Sharing Provider for PV Prosumer Clusters: A Hybrid Approach Using Stochastic Programming and Stackelberg Game. *IEEE Trans. Ind. Electron.* **2018**, *65*, 6740–6750.
17. Roth, A.E.; Erev, I. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Game Econ. Behav.* **1995**, *8*, 164–212.
18. 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.
19. Du, Y.; Li, F. Intelligent Multi-Microgrid Energy Management Based on Deep Neural Network and Model-Free Reinforcement Learning. *IEEE Trans. Smart Grid* **2020**, *11*, 1066–1076.
20. Park, L.; Jeong, S.; Kim, J.; Cho, S. Joint Geometric Unsupervised Learning and Truthful Auction for Local Energy Market. *IEEE Trans. Ind. Electron.* **2019**, *66*, 1499–1508.
21. Qian, T.; Shao, C.; Li, X.; Wang, X.; Shahidehpour, M. Enhanced Coordinated Operations of Electric Power and Transportation Networks via EV Charging Services. *IEEE Trans. Smart Grid* **2020**, *11*, 3019–3030.
22. Zhang, Y.; Zhang, Z.; Yang, Q.; An, D.; Li, D.; Li, C. EV charging bidding by multi-DQN reinforcement learning in electricity auction market. *Neurocomputing* **2020**, *397*, 404–414.
23. Kofinas, P.; Dounis, A.I.; Vouros, G.A. Fuzzy Q-Learning for multi-agent decentralized energy management in microgrids. *Appl. Energy* **2018**, *219*, 53–67.
24. Chung, J. Playing Atari with Deep Reinforcement Learning. *Comput. Ence* **2013**, *21*, 351–362.
25. Qian, T.; Shao, C.; Wang, X.; Shahidehpour, M. Deep Reinforcement Learning for EV Charging Navigation by Coordinating Smart Grid and Intelligent Transportation System. *IEEE Trans. Smart Grid* **2020**, *11*, 1714–1723.

26. Hua, H.; Qin, Y.; Hao, C.; Cao, J. Optimal energy management strategies for energy Internet via deep reinforcement learning approach. *Appl. Energy* **2019**, *239*, 598–609.
27. Kumar Nunna, H.S.V.S.; Doolla, S. Energy Management in Microgrids Using Demand Response and Distributed Storage—A Multiagent Approach. *IEEE Trans. Power Deliv.* **2013**, *28*, 939–947.
28. Xu, X.; Jia, Y.; Xu, Y.; Xu, Z.; Chai, S.; Lai, C.S. A Multi-Agent Reinforcement Learning-Based Data-Driven Method for Home Energy Management. *IEEE Trans. Smart Grid* **2020**, *11*, 3201–3211.
29. Vázquez-Canteli, J.R.; Nagy, Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Appl. Energy* **2019**, *235*, 1072–1089.
30. Lu, R.; Hong, S.H.; Zhang, X. A Dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach. *Appl. Energy* **2018**, *220*, 220–230.
31. Wan, Z.; Li, H.; He, H.; Prokhorov, D. Model-Free Real-Time EV Charging Scheduling Based on Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2019**, *10*, 5246–5257.
32. Rocchetta, R.; Bellani, L.; Compare, M.; Zio, E.; Patelli, E. A reinforcement learning framework for optimal operation and maintenance of power grids. *Appl. Energy* **2019**, *241*, 291–301.
33. Wang, H.; Huang, T.; Liao, X.; Abu-Rub, H.; Chen, G. Reinforcement Learning in Energy Trading Game among Smart Microgrids. *IEEE Trans. Ind. Electron*. **2016**, *63*, 5109–5119.
34. Xiong, R.; Cao, J.; Yu, Q. Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle. *Appl. Energy* **2018**, *211*, 538–548.
35. Yu, J.; Dou, C.; Li, X. MAS-Based Energy Management Strategies for a Hybrid Energy Generation System. *IEEE Trans. Ind. Electron.* **2016**, *63*, 3756–3764.
36. Zhang, Z.; Chong, A.; Pan, Y.; Zhang, C.; Lam, K.P. Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning. *Energy Build.* **2019**, *199*, 472–490.
37. Sun, M.; Zhao, W.; Song, G.; Nie, Z.; Han, X.; Liu, Y. DDPG-Based Decision-Making Strategy of Adaptive Cruising for Heavy Vehicles Considering Stability. *IEEE Access* **2020**, *8*, 59225–59246.
38. Xu, H.; Sun, H.; Nikovski, D.; Kitamura, S.; Mori, K.; Hashimoto, H. Deep Reinforcement Learning for Joint Bidding and Pricing of Load Serving Entity. *IEEE Trans. Smart Grid* **2019**, *10*, 6366–6375.
39. Ye, Y.; Qiu, D.; Sun, M. Deep Reinforcement Learning for Strategic Bidding in Electricity Markets. *IEEE Trans. Smart Grid* **2020**, *11*, 1343–1355.
40. Chai, B.; Chen, J.; Yang, Z.; Zhang, Y. Demand Response Management with Multiple Utility Companies: A Two-Level Game Approach. *IEEE Trans. Smart Grid* **2014**, *5*, 722–731.
41. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 1998.