energies

MDPI

Article

Multi-Agent Reinforcement Learning Approach for Residential Microgrid Energy Scheduling

Xiaohan Fang [1], Jinkuan Wang [1,*], Guanru Song [1], Yinghua Han [2], Qiang Zhao [3] and Zhiao Cao [1]

[1] College of Information Science and Engineering, Northeastern University, Shenyang 110819, China; fxh_edge@163.com (X.F.); 13194237801@163.com (G.S.); 1510384@stu.neu.edu.cn (Z.C.)
[2] School of Computer and Communication Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China; yhhan723@126.com
[3] School of Control Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China; learner_2003@163.com
* Correspondence: wjk@mail.neuq.edu.cn

Received: 25 November 2019; Accepted: 20 December 2019; Published: 25 December 2019

Abstract: Residential microgrid is widely considered as a new paradigm of the home energy management system. The complexity of Microgrid Energy Scheduling (MES) is increasing with the integration of Electric Vehicles (EVs) and Renewable Generations (RGs). Moreover, it is challenging to determine optimal scheduling strategies to guarantee the efficiency of the microgrid market and to balance all market participants' benefits. In this paper, a Multi-Agent Reinforcement Learning (MARL) approach for residential MES is proposed to promote the autonomy and fairness of microgrid market operation. First, a multi-agent based residential microgrid model including Vehicle-to-Grid (V2G) and RGs is constructed and an auction-based microgrid market is built. Then, distinguish from Single-Agent Reinforcement Learning (SARL), MARL can achieve distributed autonomous learning for each agent and realize the equilibrium of all agents' benefits, therefore, we formulate an equilibrium-based MARL framework according to each participant' market orientation. Finally, to guarantee the fairness and privacy of the MARL process, we proposed an improved optimal Equilibrium Selection-MARL (ES-MARL) algorithm based on two mechanisms, private negotiation and maximum average reward. Simulation results demonstrate the overall performance and efficiency of proposed MARL are superior to that of SARL. Besides, it is verified that the improved ES-MARL can get higher average profit to balance all agents.

Keywords: residential microgrid; energy scheduling; vehicle-to-grid; multi-agent reinforcement learning; game theory; equilibrium selection

1. Introduction

1.1. Motivation

In recent years, a microgrid-based family energy framework has increasingly attracted attention. This emerging residential energy system contains distributed Renewable Generations (RGs), household load appliances, and Energy Storage Units (ESUs). The application of residential microgrid reduces the user's dependence on the main grid and improves the autonomy and flexibility of the family power system [1]. In the practical field, resident users, RGs and ESUs constitute a small and independent microgrid market [2,3]. It's essential to formulate an intelligent and effective residential Microgrid Energy Scheduling (MES) mechanism for coordinating and balancing the benefits of all members, meanwhile, guaranteeing the members' self-decision ability and information security.

Besides, Electric Vehicles (EVs) are becoming more and more widespread in resident life due to its various advantages. EVs can effectively solve the pollution problem from traditional energy; in addition, the energy consumption cost of EVs is cheaper than gasoline vehicles [4,5]. Moreover, Vehicle-to-Grid (V2G) mode allows EVs to discharge energy to the power grid for participating in market scheduling and earning profit [6,7].

Considering the integration of EVs and RGs, residential MES becomes more complicated due to high uncertainties from EVs and RGs, such as RGs output and EVs' usage attributes [8,9]. Therefore, some challenges exist in residential MES for previous studies shown in Section 1.2. First, it's difficult to build a precise scheduling model that can attend to all the uncertain parameters. Second, the common centralized dispatch methods need a completely open market environment and could be problematic with a large system and multiple constraints. Third, most works are myopic solutions considering the current objectives, instead of a long-term optimization.

As a solution to confront these challenges, a model-free machine learning method, Reinforcement Learning (RL) has got a good effect on complicated decision-making problems. However, most RL methods adopt Single-Agent RL (SARL) to obtain the optimal policy, some defects exist in SARL whether the algorithm mode is centralized or decentralized. In centralized SARL, first, microgrid control center gathers all participants' information and implements RL tasks centrally, each participant passively performs learning results from control center, rather than learning autonomously; second, participants need to upload necessary information to control center, thus part of private information may be at risk of leaking; third, the penetration of EVs and RGs will increase computational complexity of SARL, even lead to the curse of dimensionality. On the other hand, in decentralized SARL, each participant can learn and make decisions according to individual information and environment, but SARL is a kind of selfish learning to maximize self-reward, instead of considering overall profit and supply-demand balance of microgrid [10]; besides, the learning process of each participant is based on itself's available information, they can't obtain other agents' some confidential information such as providers' quotes and demand response behaviors of users, therefore, the agents' learning processes are imprecise and there are no interactions among agents.

To address all the above issues, this paper presents a Multi-Agent RL (MARL) approach for residential MES. MARL is a distributed RL in multi-agent environment that can be seemed as a combine of RL and game theory [11]. Although the MARL framework is applicable to residential MES to construct a distributed microgrid RL architecture, some limitations restrict the application of existing MARL methods in the microgrid field. First, most MARL algorithms require agents to replicate other agents' value functions and to calculate the equilibria for all joint-actions, which are computationally expensive. Then, if agents' information is not be fully shared (incomplete information game), it's difficult to obtain a definite equilibrium solution [12]. Finally, the solved equilibrium solutions may not be unique, how to select the optimal equilibrium to balance all agents' rewards and to ensure the convergence of MARL are noteworthy [11,13]. In this paper, we present an Equilibrium Selection-based MARL (ES-MARL) method, an optimal Equilibrium Selection (ES) is adopted according to two mechanisms, that is private negotiation with each agent and maximum average benefit method.

1.2. Related Works

Several studies about residential MES considering EVs or RGs have been published using various approaches [14–20]. For example, a virtual power plant based on linear programming was used as a combination of wind generators and EVs to schedule the market in [14]. In [15], a dynamic programming-based economic dispatch for community microgrid is formulated. The authors in [16] proposed a hierarchical control method to achieve coordination scheduling integrating EVs and wind power. In [17], an EV coordination management algorithm was presented to minimize the load variance and to enhance the distribution network's efficiency. A game theory-based retail microgrid market was built and a Nikaido-Isoda Relaxation approach was adopted to get the optimal solution [18]. A hierarchical control framework for microgrid energy management system with RGs and an ESU is

proposed [19]. Ref. [20] studies about a day-ahead scheduling of integrated home-type microgrid and adopts a mixed-integer linear programming algorithm to achieve optimal energy management.

These previous studies cannot take all challenges in MES into account. Therefore, RL-based method is adopted as solution. RL approach learns the optimal policies through trial-and-error mechanism, that does not depend on the prior knowledge of system model information, and RL has been widely used in energy scheduling of smart grid and EVs [12,21–23]. For instance, ref. [12] focused on the smart grid market based on double auction, an adaptive RL was used to find the Nash Equilibrium (NE) of energy trading game with incomplete information. The authors in [21] proposed an RL-based dynamic pricing and energy consumption scheduling algorithm for the microgrid system. In [22], a batch RL approach was adopted in residential demand response to make a day-ahead consumption plan. In [23], authors raised RL-based real-time power management to solve the power allocation for hybrid energy storage system in a plug-in hybrid EV.

Moreover, in this paper, a MARL method is used for sequential decision making in multi-agent environment where traditional SARL is difficult to deal with. MARL has been adopted in some fields, such as vehicle routing problem [24] and thermostatically loads modeling [25]. The most universal MARL is equilibrium-based MARL, whose framework accords with Markov games and the evaluation of the learning process is based on all agents' joint behaviors, the equilibrium concept from game theory is introduced to denote optimal joint action [26–30]. The earliest proposed equilibrium-based MARL was the Minimax-Q algorithm which considered two agents' zero-sum game, two agents try to maximize and minimize one reward function [26]. Authors in [27] proposed Nash-Q algorithm for non-cooperative general-sum Markov games, and the NE solution was adopted to define value function. In [28], a friend-or-foe Q-learning algorithm was presented for obtaining different solutions based on agents' relationships. In [29], Correlated-Q learning was proposed base on correlated equilibrium solution, which allows for the possibility of dependencies in the agents' optimization. In [30], authors introduced a "Win or Learn Fast" (WoLF) mechanism to form a variable learning rate based MARL method.

These papers have made contributions on the domain of microgrid energy scheduling or RL algorithm. Through the analysis and improvement of these studies, in this paper, a MARL approach is adopted for management and decision of distributed microgrid market.

1.3. Contributions

To sum up, the principal contributions of this paper are summarized as follows.

- A framework for residential MES with V2G system is built. All participants in the microgrid and the auction-based microgrid market mechanism are modeled.
- MARL algorithm is introduced for the first time into a residential microgrid. The RL model (states, actions, and immediate reward) of each agent is formulated.
- An improved ES-MARL is proposed, the Equilibrium Selection Agent (ESA) calculates the corresponding equilibrium solution by negotiating with agents and selects the optimal solution based on maximum average reward.

1.4. Organization

The structure of this paper is as follows. In Section 2, we present the microgrid model and market mechanism. In Section 3, we summarize the MARL theory and the microgrid agents' design of MARL. In Section 4, we propose the ES-based MARL method. Section 5 presents the simulation results and analyses. We conclude the paper in Section 6.

2. System Model

2.1. Residential Microgrid Description

With the development of RGs and EVs, the residential microgrid has been increasingly used. For a urban residential area, the fluctuations of daily load curves are high and the distributions of residents, RGs and EVs connecting are stochastic; moreover, participants have higher requests for autonomy energy management from the perspective of economy and privacy protection; besides, residential microgrid should consider more about environmental concerts and power safety. In this paper, a distributed framework for residential microgrid which is more applicative to meet above requirements is adopted. As depicted in Figure 1, the 9-node residential microgrid is built based on multi-agent system [31,32], all participants are modeled as profit-aware intelligent agents with abilities of autonomous learning and decision-making; agents in the market should comply with microgrid market mechanism and follow market scheduling based on global optimization goal.

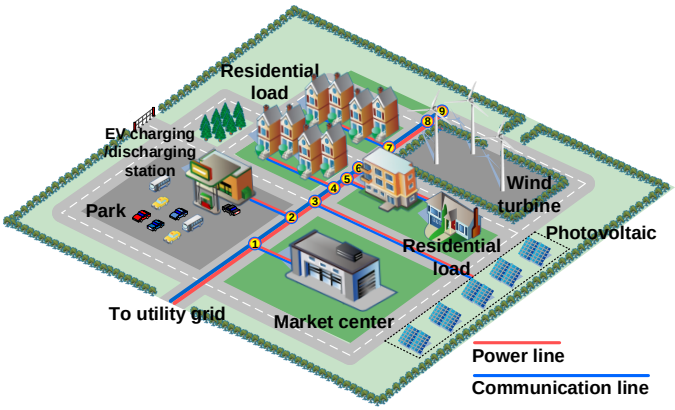

Figure 1. System model of 9-node residential microgrid.

Based on the roles in the market, microgrid agents are divided into different clusters: RGs are independent supplier agents; users are consumer agents and can join in microgrid scheduling through demand response; a unified management for EVs acts as supplier agents or consumer agents according to overall charging/discharging action; besides, manager agents (e.g., market operator and equilibrium selector) are set to maintain market operation. The primary objective for the microgrid in grid-connected mode is to achieve maximum autonomy, that is to provide the necessary load demand with the minimum dependency on Utility Grid (UG). Besides, agents' benefits should be considered to realize a unanimously acceptable balance. All agents' concrete models are described as follows.

2.2. Microgrid Components

2.2.1. Electric Vehicles

EVs are both power consumers and power suppliers in the microgrid market. Considering the negative effects of V2G, extra battery degradation cost and the impact on subsequent travel should be considered to evaluate the net profit of V2G. Distinct from stationary ESUs, EVs' charging/discharging actions will be affected by owners' traveling habits and stochastic behaviors of EVs usage. Besides, specific operational and technical constraints of EVs should be noticed.

(1) EV Travel Behaviors and Constraints: EVs charging/discharging scheduling should take travel demand as premise. The customary travel habits (e.g., arrival time, departure time and travel distance) follow a similar pattern based on the owner's intentions, besides, the random characters of travel behaviors are considerable [33].

The State-Of-Charge (SOC) of EV i at slot t is defined as soc_t^i, n_e is the current number of EVs, $i \in [1, 2, 3, \cdots, n_e]$. The constraints of EVs SOC is:

$$soc_{min} \leq soc_t^i \leq soc_{max} \tag{1}$$

where soc_{min} and soc_{max} are minimum and maximum limits of EV battery SOC, respectively. The charging/discharging limitations are:

$$-v_{max}^{id} \leq v_t^i \leq v_{max}^{ic} \tag{2}$$

where v_t^i is EV i' charging/discharging power at slot t, v_t^i is positive when EV is charging and is negative when EV is discharging. $v_{max}^{id} = r_d^i t$ and $v_{max}^{ic} = r_c^i t$, where r_c^i and r_d^i are EV i' battery charging and discharging rate. Then, we have:

$$soc_{t+1}^i = soc_t^i + \frac{v_t^i}{b_m^i} \tag{3}$$

where b_m^i is EV i' battery capacity. If EV i will leave at slot t, SOC should satisfy the travel demand as:

$$soc_t^i \geq soc_{min} + soc_{dis}^i \tag{4}$$

where soc_{dis}^i is EV i' minimum SOC for departure. Remarkably, only if EVs adopting slow-charge can participate in the microgrid market; if EVs need urgent travel, the fast-charge will be chosen. The EV owner's travel habits are from data statistics, arrival time and departure time follow normal distribution; and travel distance follows a log-normal distributed.

(2) EV Battery Degradation Cost: The life of EV battery declines along with repeating charging/discharging cycles. For lithium iron phosphate battery adopted in this paper, low temperature weaken performance and high temperature curtails battery life. The use of battery in moist condition should be avoided. Besides, a prolonged period of overvoltage during long travel may stress the battery [34]. In sum, we consider that EVs battery degradation depends on the number of cycles. According to [35], the battery degradation cost function can be approximately expressed as:

$$C^{v_t^i} = \frac{k}{100} \frac{|v_t^i|}{b_m^i} C_b^i \tag{5}$$

where $C^{v_t^i}$ is battery degradation cost of v_t^i, C_b^i represents battery cost, k is the slope of the linear approximation of the battery life as a function of the cycles.

(3) EV Aggregator and EV Secondary Scheduling: In our model, an EV Aggregator (EVA) is adopted to manage all EVs' participation in the market. The introduction of EVA facilitates the model extendibility and adjustment when the number of EVs changes, besides, the reduction of agents number can improve the convergence speed of learning algorithm.

Base on [33], all local EVs participating in the microgrid market (primary market) form a secondary scheduling system of EVs, EVA is the manager of the secondary scheduling system. At the beginning of each slot, EVA arranges each EV's charging/discharging amount based on the optimal global charging/discharging action from the primary market, besides, some rules should be obeyed as follows.

- EVs' travel demand should be satisfied, first of all, EVA considers EVs' travel demand two hours ahead of departure and guarantees SOC is more than soc_{dis}^i.
- EVA arranges EVs charging/discharging sequence according to SOC level, if $soc_t^i < soc_{min}^{cha}$, this EV can't discharge and should be arranged to charge, where soc_{min}^{cha} is charge warning limit.

- The total charge/discharge amount from the primary market is the scheduling objective of the secondary system, the sum of EVs' charge/discharge should be equal to the total amount.

2.2.2. Users' Loads

The residential appliances keep approximately steady in identical period during the same season, therefore, users' load demand can be predicted accurately. User i's load demand at slot t is written as d_t^i, $d_t^i \in [d_{t,min}^i, d_{t,max}^i]$, $d_{t,min}^i$ and $d_{t,max}^i$ are minimal essential demand and maximal available demand, respectively. n_u is the number of users, $i \in [1, 2, 3, \cdots, n_u]$. According to operating profiles, the household load can be categorized into three types as follows [36].

(1) fixed loads: this kind of load demand can't be changed to guarantee the devices in working order, such as web servers and medical instruments. User i' critical loads profile is denoted as $d_t^{i,f}$.

(2) curtailable loads: the demand can be cut down for reducing consumption, such as heating or cooling devices. $d_t^{i,c}$ denotes the load profile of curtailable loads, and we denote the $0 \leq l_t^{i,c} \leq 1$ as the ratio of load curtailment. Then, we have:

$$d_{t,min}^{i,c} \leq (1 - l_t^{i,c}) d_t^{i,c} \leq d_t^{i,c} \tag{6}$$

where $d_{t,min}^{i,c}$ is the fixed part of curtailable loads, if $d_{t,min}^{i,c} = d_t^{i,c}$, this kind of devices becomes fixed loads; if $d_{t,min}^{i,c} = 0$, user can turn off these devices.

(3) shiftable loads: the operation period can be postponed to avoid the load peak, such as the washing machine. The shiftable loads profile is $d_t^{i,s}$, if user postpones the loads demand $0 < l_t^{i,s} \leq d_t^{i,s} + l_{t-1}^{i,s}$, the shiftable part $l_t^{i,s}$ will defer to the next time slot for consideration.

In sum, user i' total load demand $d_t^i = d_t^{i,f} + d_t^{i,c} + d_t^{i,s} + l_{t-1}^{i,s}$. Therefore, after demand response, user i' actual load demand can be denoted as:

$$l_t^i = d_t^{i,f} + (1 - l_t^{i,c}) d_t^{i,c} + d_t^{i,s} + l_{t-1}^{i,s} - l_t^{i,s} \tag{7}$$

Users can regulate the schedules of curtailable loads and shiftable loads to control the load consumption; meanwhile, load curtailment and shift will reduce the user's consumption satisfaction. In this paper, a utility function $U(l_i^t)$ which represents user's satisfaction is adopted as:

$$U(l_i^t) = \begin{cases} \omega l_i^t - \frac{\beta}{2}(l_i^t)^2 & 0 \leq l_i^t \leq \frac{\omega}{\beta} \\ (\omega)^2 / 2\beta & l_i^t > \frac{\omega}{\beta} \end{cases} \tag{8}$$

where ω indicates users' action (a larger ω implies a larger utility); β denotes the utility saturation. The utility function should be non-decreasing and concave. Similarly, in the learning process of MARL, a User Aggregator (UA) is introduced to centrally arrange the demand responses of all users.

2.2.3. Renewable Generations

RGs' generation capacity can be derived from accurate short-term forecasts via historical data and environmental data. The random characteristics of RGs' generation are represented as stochastic normal distribution based on forecast results.

From [37], the generation of PV g_t^{pv} is closely affected by the weather factors, g_t^{pv} is determined as:

$$g_t^{pv} = \eta \times S_{pv} \times I_t [1 - 0.005(T - 25)] \tag{9}$$

where η is conversion rate of PV array (%); S_{pv} is area of PV array (m^2); I_t is solar irradiance which is from the Beta probability density function (kW/m^2) [1]; T is the average temperature during t (°C).

The main element influencing the output of WT is wind speed V_t. The WT generation g_t^{wt} is a piecewise function of V_t denote as below [38]:

$$g_t^{wt} = \begin{cases} 0 & V_t < V_{in};\ V_t \geq V_{off} \\ (V_t - V_{in})/(V_r - V_{in}) \times P_r^{wt} & V_{in} \leq V_t \leq V_r \\ P_r^{wt} & V_r \leq V_t < V_{off} \end{cases} \tag{10}$$

where V_{in} and V_{off} are the cut-in wind speed and cut-off wind speed, respectively; V_r is the rated wind speed; P_r^{wt} is the rated power of WT.

2.3. Transaction with Utility Grid

Microgrid keeps supply-demand balance through energy trade with UG, the transaction prices between UG and microgrid are fixed by UG. Here, UG adopts a real-time price mode that the price is variable at each time slot. The trade price between UG and microgrid market is summarized as a tuple $p_t^u = (p_t^{u,p}, p_t^{u,s})$, $p_t^{u,p}$ is purchase price from UG, and $p_t^{u,s}$ is sell price to UG. To avoid energy arbitrage, $p_t^{u,s}$ is perpetually lesser than $p_t^{u,p}$, a factor $\rho = p_t^{u,s}/p_t^{u,p}$ is defined as sell/purchase price ratio [39].

2.4. Microgrid Market

As suggested by [33,40], a real-time microgrid market is constructed based on a one-side dynamic bidding model. In the microgrid market, supplier agents provide their quotes with bid price and supply amount; consumer agents submit their energy demand which can be optimized by adjusting their consumption behaviors. The above information of sellers and consumers aggregate in a non-profit Market Operator Agent (MOA), whose duty is to make the market clearing price p_t^m and to calculate each agent's energy sell/purchase amount. The microgrid market operates per time slot t, abiding the following principles.

- In the market clearing process, MOA sorts sellers' quotes in increasing order of prices, then the demand will be matched according to the ranking and respective supply. The bid price of the last adopted quote is defined as the marginal price, that is the market clearing price p_t^m.
- If the energy supply is lacking to support the load demand, MOA will purchase energy from UG. The purchase price from UG is higher than the sellers' bids. The additional expenditure will be charged averagely by all purchasers.
- If the supply exceeds the demand, the surplus energy will be sold to UG. If more than one seller offers the market clearing price, their energy sold to microgrid and UG are arranged based on the same proportion of their supply.

3. MARL Method for Microgrid Market

For a multi-agent based microgrid system, the most important task is to generate agents' distributed strategies to schedule their behaviors in the market. Moreover, it's significant to ensure all agents' benefits balance. In this section, a MARL method is introduced to solve this issue about how to generate and coordinate the autonomous strategies for all agents.

3.1. Overview of MARL

RL algorithm is an unsupervised machine learning method for sequential decision problems. In SARL, agents interact with the environment by executing actions, the environment then feeds back an immediate reward to evaluate the selected action. Transfer to MARL, the relationships with both cooperations and competitions exist among agents, and agents' rewards are influenced by other agents' states and actions. SARL is built on the framework of Markov Decision Process (MDP), but in MARL, the framework is Markov game, which is the combination of MDP and game theory [41,42].

Definition 1. *Markov Game. An n-agent ($n \geq 2$) Markov game is a tuple $\langle S, A^1, \ldots, A^n, r^1, \ldots, r^n, P \rangle$, where n is the number of autonomous agents; S is the state space of system and environment; A^i is the action space of agent i, $i \in [1, 2, \ldots, n]$, then the joint-action space is defined as $A = A^1 \times \ldots \times A^n$ and a joint-action $\vec{a} = (a^1, a^2, \ldots, a^n)$; r^i is the immediate reward function of agent i; P is the transition function, denotes the probability distribution when an action is executed, the current state transfers to a new state, $P : S \times A \times S \to [0, 1]$.*

As shown in Figure 2, compared with SARL, the main distinction of MARL is that the reward function and state transition for agents are based on the joint-action $\vec{a}$. In a pure strategy game, the agents' joint-action is defined as $\vec{a} = (a^1, a^2, \ldots, a^n)$, in MARL, when $\vec{a}$ is applied and the state transfers from s to s', an action evaluation Q-function for agent i, $Q^i(s, \vec{a})$ is defined as:

$$Q^i(s, \vec{a}) = r^i(s, \vec{a}) + \gamma \sum_{s' \in S} P_{ss'}(\vec{a}) V^i(s') \tag{11}$$

where $V^i(s')$ is the maximum discounted cumulative future reward starting from state s'; $0 \leq \gamma < 1$ is the discount factor, which indicates the weight of future reward.

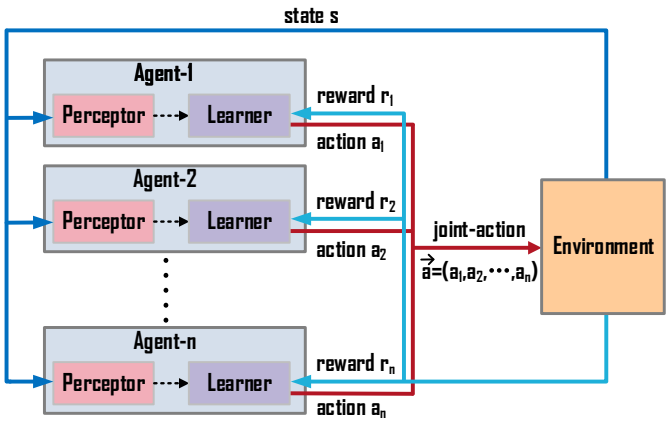

Figure 2. Principle diagram of MARL.

In SARL, agent i's goal is to find an optimal action policy $a^i_*(s)$ to maximize Q-function, but the goal in MARL is to find an optimal joint-action $\vec{a}_*(s)$ to coordinate all the Q-functions $\{Q^i(s, \vec{a})\}_{i=1}^n$ to a global equilibrium. A concept of equilibrium solution from game theory is introduced into MARL. The generally used equilibrium solution is NE [27].

Definition 2. *Nash Equilibrium. In a pure strategy Markov game, a NE solution is defined as a joint-action $\vec{a}_* = (a^1_*, \ldots, a^i_*, \ldots, a^n_*)$, satisfying the following criterion for all n-agents:*

$$Q^i(\vec{a}_*) \geq \max_{a^i \in A^i} Q^i(a^1_*, \ldots, a^{i-1}_*, a^i, a^{i+1}_*, \ldots, a^n_*) \tag{12}$$

An intuitive comprehension about NE is that, for all agents, if other agents don't change their actions a^{-i}, agent i can't improve its utility function $Q^i(\vec{a})$ by changing itself's action a^i, where a^{-i} is the joint-action of all agents except agent i. At time slot t, the iterative modified formula of Nash-Q MARL is expressed as:

$$Q^i_{t+1}(s, \vec{a}) = (1 - \alpha) Q^i_t(s, \vec{a}) + \alpha [r^i_t(s, \vec{a}) + \gamma Nash Q^i_t(s')] \tag{13}$$

where $0 < \alpha < 1$ is the learning rate, whose value decides convergence speed of MARL. $Nash Q^i_t(s')$ is agent i' Q-value with the selected NE in the next state s'.

In general, the major structures of other equilibrium-based MARL algorithms are similar to Nash-Q MARL, the main difference is various selected equilibrium in the learning process.

3.2. Agents Design for Microgrid MARL

The equilibrium-based MARL is applicable in a mixed task (competitions coexist with cooperations) multi-agent system, in the residential microgrid market, there are both competitive relations (e.g., suppliers' quotes) and cooperative relations (e.g., sellers and purchasers collaborate to achieve supply-demand balance and microgrid autonomy). Detailed formulations of MARL used in residential MES are introduced in this section.

First of all, all agents' MDP models are designed as follows.

3.2.1. EV Aggregator Agent

In the primary microgrid market, EVA agent participates in the market as a centralized agent of all EVs. EVA decides the total charging/discharging in the market based on MARL results. In EVA's learning process, the current local EVs number, SOC and travel demand of EVs should be considered.

State: The state-base for EVA agent at slot t is defined as:

$$s_t^{eva} = (t, soc_t, nev_t) \tag{14}$$

where soc_t is the average SOC of local EVs; nev_t is the number of local EVs, which connect with microgrid. According to t, we can get the UG price p_t^u.

Action: EVA's actions include total charging/discharging power v_t and EVA's quote p_t^{eva}. $v_t > 0$, EVA acts as a purchaser; $v_t < 0$, EVA serves as a seller; $v_t = 0$ means that there is no energy trade. Only when $v_t < 0$, p_t^{eva} is existent. The action-base of EVA agent is denoted as:

$$a_t^{eva} = (v_t, p_t^{eva}) \tag{15}$$

Considering that only if the EV is connected to microgrid (non-movement state), the charge/discharge action can be executed [43], so the action of EVA is constrained by average SOC soc_t, local EVs' number nev_t and travel demand (denoted by soc_{dis}). The total charging/discharging power v_t is confined as:

$$min\{v_{max}^{id}\}_{i=1}^{n_e} \le v_t \le min\{v_{max}^{ic}\}_{i=1}^{n_e} \tag{16}$$

Besides, when $soc_t \le soc_{min} + min\{v_{max}^{id}/b_m^i\}_{i=1}^{n_e}$, EVA can only select energy purchase actions; when $soc_t \ge soc_{max} - max\{v_{max}^{ic}/b_m^i\}_{i=1}^{n_e}$, EVA can only select energy sell actions.

Reward: According to statistics, more than 90% of EV users will charge the SOC up to 60% before leaving. The user's anxiety due to the worry about exhausting energy on the road aggravates increasingly with the decline of SOC. Therefore, the immediate reward function of EVA should combine economy, battery degradation, and user's anxiety, which is defined as:

$$\begin{cases} r_t^{eva} = \nu_c[(soc_{max} - soc_t)\overline{b_m} + v_t]^2 - p_t^m v_t^m - p_t^{u,p} v_t^u - C_t^{v_t} & v_t \ge 0,\ EV\ charges \\ r_t^{eva} = p_t^m v_t^m + p_t^{u,s} v_t^u - C_t^{v_t} - \nu_d[(soc_{max} - soc_t)\overline{b_m} - v_t]^2 & v_t < 0,\ EV\ discharges \end{cases} \tag{17}$$

where ν_c and ν_d are the charging/discharging anxiety coefficients. $\overline{b_m}$ is the average battery capacity of all EVs. v_t^m and v_t^u are the energy trade with microgrid and with UG, respectively; when $v_t \ge 0$, $v_t^m + v_t^u = v_t$; when $v_t < 0$, $v_t^m + v_t^u = -v_t$.

3.2.2. User Aggregator Agent

From Section 2.2.2, users can control curtailable loads and shiftable loads to reduce cost; meanwhile, the load adjustments depress users' satisfaction. UA agent takes a uniform demand

response action based on the results of MARL, then the demand response task is assigned averagely to all users in the secondary scheduling.

State: The state-base of UA agent at slot t is defined as:

$$s_t^{ua} = (t, d_t) \tag{18}$$

where t is current time slot; d_t is cumulative load demand of all users without demand response.

Action: The action-base of UA agent at slot t is defined as:

$$a_t^{ua} = (l_t^c, l_t^s) \tag{19}$$

where l_t^c is the total ratio of load demand curtailment; l_t^s is the cumulative shiftable load demand. Then, $d_t = d_t^f + d_t^c + d_t^s + l_{t-1}^s$, where d_t^f, d_t^c and d_t^s are total demand of fixed load, curtailable loads and shiftable loads, respectively.

Reward: Similar to Equation (7), the actual cumulative load demand is defined as

$$l_t = d_t^f + (1 - l_t^c)d_t^c + d_t^s + l_{t-1}^s - l_t^s \tag{20}$$

UA agent's immediate reward function r_t^{ua} is defined as the difference between users' total utility function and energy purchase expenses.

$$r_t^{ua} = U(l_t) - p_t^m l_t^m - p_t^{u,p} l_t^u \tag{21}$$

where $l_t^m + l_t^u = l_t$, l_t^m and l_t^u are the energy purchase from microgrid market and UG, respectively.

3.2.3. RG Agents

There are two kinds of RGs, Photo-Voltaic (PV) and Wind Turbine (WT), the output of RGs is based on the short-term forecasts, and random distributions with forecast values will be adopted to embody the uncertainties of RGs' generation. All of the RGs' generation will be put into the market at the current time slot.

State: The state of RG agents is current time slot t.

$$s_t^{pv} / s_t^{wt} = t \tag{22}$$

Action: The actions of RG agents are denoted as:

$$a_t^{pv} = p_t^{pv} \tag{23}$$

$$a_t^{wt} = p_t^{wt} \tag{24}$$

where p_t^{pv} / p_t^{wt} are the quote prices of PV agent and WT agent.

Reward: The RG agents' immediate reward functions are profit functions as:

$$r_t^{pv} = p_t^m g_t^{pv,m} + p_t^{u,s} g_t^{pv,u} - C^{pv}(g_t^{pv}) \tag{25}$$

$$r_t^{wt} = p_t^m g_t^{wt,m} + p_t^{u,s} g_t^{wt,u} - C^{wt}(g_t^{wt}) \tag{26}$$

where for the tuple rg=(pv/wt), $g_t^{rg,m}$ is the portion sold to microgrid; $g_t^{rg,u}$ is the portion sold to UG, $g_t^{rg,m}$+$g_t^{rg,u}$=g_t^{rg}. C^{rg} is the generation cost function, which is considered as a quadratic function as:

$$C^{rg}(g_t^{rg}) = c_1(g_t^{rg})^2 + c_2 g_t^{rg} + c_3 \tag{27}$$

where c_1, c_2, c_3 are pre-determined parameters which are different for PV and WT, and $c_1, c_2, c_3 \geq 0$.

3.3. MARL Method for Residential MES

Based on agents' MDP model designs, the equilibrium-based MARL is adopted. The general framework of agents' learning process for microgrid MARL is shown as Algorithm 1.

At line 3 of Algorithm 1, ϵ-greedy policy denotes that the agent selects a random action with a probability of $1-\epsilon$ and selects a joint action which makes system achieve equilibrium with a probability of ϵ. $\Phi_t^i(s_{t+1}^i)$ is the expected value of the equilibrium in state s_{t+1}^i.

Algorithm 1 General Framework of Microgrid MARL

Input:

Agent set N; state space S^i; action space A^i; learning rate α^i; discount factor γ^i; joint action space A.

1: Initialize $Q^i(s^i, \vec{a}) \leftarrow 0$; initialize state $s_t^i \in S^i$, and $t = 0$;
2: **repeat**
3: For each time slot t, each agent selects $a_t^i \in A_i$ with ϵ-greedy policy to form a joint-action $\vec{a}_t \in A$;
4: MOA calculates the market clearing price p_t^m and the energy should be traded of each agent;
5: Each agent obtains the experience $(s_t^i, \vec{a}_t, r_t^i, s_{t+1}^i)$;
6: Each agent updates the Q-matrix $Q_t^i(s_t^i, \vec{a}_t)$:
$Q_t^i(s_t^i, \vec{a}_t) \leftarrow (1-\alpha^i)Q_t^i(s_t^i, \vec{a}_t) + \alpha^i[r_t^i(s_t^i, \vec{a}_t) + \gamma^i \Phi_t^i(s_{t+1}^i)]$
7: $s_t^i \leftarrow s_{t+1}^i$, $t \leftarrow t+1$;
8: **until** Q-matrix $Q^i(s^i, \vec{a})$ converges.

Output:

The optimal Q-matrix $Q^i(s^i, \vec{a})$ for each agent.

4. Improved Equilibrium-Selection-Based MARL

As mentioned in Section 1.1, some limitations existing in common MARL algorithms, are summarized as follows.

- In the learning process of common MARL, agent updates and saves other agents' value functions in each iteration step, that will cause a huge computation work, even in a small scale environment with two or three agents.
- In order to work out the equilibrium solution, MARL needs agents to share their states, actions and value functions, including some privacy information, which is unrealistic in some situations.
- In each learning iteration step, there is perhaps more than one equilibrium solution, different equilibria bring different updates of value functions, which may lead to non-convergence of the algorithm. Besides, it's hard to ensure fairness for selecting an equitable equilibrium because agents' rewards are different with different equilibria.

Therefore, we present an ES-MARL algorithm to address these issues. We set up an Equilibrium Selection Agent (ESA), whose function is to separately negotiate with all agents to get the equilibria solution set and to select the optimal equilibrium based on the maximum average benefit method.

4.1. Negotiation for Equilibrium Solution Set

In an incomplete information game, agent's reward information is incompletely public to other agents, agents can't obtain other agents' value functions to compute the equilibria. To solve this problem, according to [11], ESA is adopted as a neutral negotiator to communicate with each agent privately to obtain their potential equilibrium set following these steps:

1. At the beginning of slot t, agent i finds its potential NE set Z_{ne}^i and sends it to ESA (concrete steps for finding potential NE set are shown in Algorithm 2).
2. ESA selects the joint-action $\vec{a}^j \in A$, which meets the criteria $\forall Z_{ne}^i, \vec{a}^j \in Z_{ne}^i$, into the final equilibrium set Z_e, the selected $\vec{a}^j$ is the pure strategy NE solution of game $Q^1(s_t), \ldots, Q^n(s_t)$.

3. If there is no joint-action satisfying NE, ESA selects $\vec{a}^j$ whose number of satisfying $\vec{a}^j \in Z_{ne}^i$ k^j, is the most and $k^j > 0.5n$ (more than half agents get to equilibrium), then adds $\vec{a}^j$ into Z_e.

Now, the equilibria set Z_e is the candidate set by negotiations, the element number of Z_e may be more than one.

Algorithm 2 Equilibrium Selection Process of MARL

Input:

Agent set N; current state $s^i \in S^i$; joint-action space A; agent number n; weight factor β.

Potential NE set $\{Z_{ne}^i\}_{i=1}^n \leftarrow \emptyset$; final equilibrium set $Z_e \leftarrow \emptyset$;
for each $\vec{a}^{-i} \in A^{-i}$ **do**
  $a^i \leftarrow \arg\max_{a^i} Q^i(s^i, a^i, \vec{a}^{-i})$;
  $Z_{ne}^i \leftarrow Z_{ne}^i \cup \{(a^i, \vec{a}^{-i})\}$;
end for
Each agent sends its Z_{ne}^i to ESA;
$Z_{ne}^{sum} = Z_{ne}^1 \cup Z_{ne}^2 \ldots \cup Z_{ne}^i$;
for each $\vec{a}^j \in Z_{ne}^{sum}$ **do**
  if $\forall Z_{ne}^i, \vec{a}^j \in Z_{ne}^i$ **then**
    $Z_e \leftarrow Z_e \cup \{\vec{a}^j\}$;
  else
    Calculate k^j, which is the number of $\vec{a}^j \in Z_{ne}^i$;
  end if
  if $Z_e \neq \emptyset$ **then**
    Z_e is the final pure strategy NE set;
  else
    if $k^j > 0.5n$ **then**
      $\vec{a}^j \leftarrow \arg\max_{\vec{a}^j} k^j$;
      $Z_e \leftarrow Z_e \cup \{\vec{a}^j\}$, Z_e is the final suboptimal set;
    else
      $Z_e \leftarrow A$;
    end if
  end if
end for
for each $\vec{a}^j \in Z_e$ **do**
  Calculate the value of joint benefit function:
  $J = \frac{1}{n} \sum_{i=1}^{n} Q^i(s^i, \vec{a}^j)$;
end for
$\vec{a}_* \leftarrow \arg\max_{\vec{a}^j} J$;

Output:

The optimal equilibrium $\vec{a}_*$.

4.2. Equilibrium Selection Based on Maximum Average Reward

If the element number of Z_e is more than one, the update of Q-function shown as Equation (13) will get different values based on different selected equilibria. In this paper, a maximum average reward method is adopted to help ESA selecting the optimal equilibrium to guarantee algorithmic efficiency and fairness.

Here, we introduce an average reward function J which denotes all agents' average reward as:

$$J = \frac{1}{n} \sum_{i=1}^{n} Q^i(s^i, \vec{a}^j) \tag{28}$$

ESA selects the optimal equilibrium (joint action $\vec{a}_*$), whose corresponding value of J is maximum. The selected optimal equilibrium is used in MARL to update Q-functions. The ES process is shown in Algorithm 2.

Based on the optimal equilibrium joint-action $\vec{a}_*$, an improved ES-MARL algorithm is shown in Algorithm 3. Private negotiations between ESA and other agents can avoid redundant updates of Q-functions for each agent and protect the privacy information; meanwhile, the optimal ES process can combine all agents' benefit and promote global welfare. The fairness, safety, and efficiency of the microgrid market are guaranteed with ES-MARL.

Algorithm 3 ES-Based MARL Algorithm

Input:

Agent set N; state space S^i; action space A^i; learning rate α^i; discount factor γ^i; joint action space A.

1: Initialize $Q^i(s^i, \vec{a}) \leftarrow 0$; initialize state $s_t^i \in S^i$, and $t = 0$;
2: **repeat**
3: For each time slot t, agents select random $a_t^i \in A_i$ with probability ϵ or adopt an optimal equilibrium joint-action $\vec{a}_t^*$ based on Algorithm 2 with probability $1 - \epsilon$;
4: MOA calculates the market clearing price p_t^m and the energy should be traded of each agent;
5: Each agent obtains the experience $(s_t^i, \vec{a}_t, r_t^i, s_{t+1}^i)$;
6: ESA computes the optimal equilibrium of next slot $\vec{a}_{t+1}^*$ based on Algorithm 2;
7: Each agent updates $Q_t^i(s_t^i, \vec{a})$ (with ϵ-greedy policy):
$Q_t^i(s_t^i, \vec{a}_t) \leftarrow (1 - \alpha^i) Q_t^i(s_t^i, \vec{a}_t) + \alpha^i [r_t^i(s_t^i, \vec{a}_t) + \gamma^i Q_t^i(s_{t+1}^i, \vec{a}_{t+1}^*)]$
8: $s_t^i \leftarrow s_{t+1}^i$, $t \leftarrow t + 1$;
9: **until** Q-matrix $Q^i(s^i, \vec{a})$ converges.

Output:

The optimal Q-matrix $Q^i(s^i, \vec{a})$ for each agent.

4.3. Overall Process of Proposed ES-MARL Approach

To sum up, we present a flowchart shown in Figure 3 about proposed ES-MARL approach for microgrid energy scheduling, including the MARL training process and MARL application process. From Figure 3, we can see in the training process of ES-MARL algorithm, the ES procedure (shown in Section 4, Algorithm 3) is responsible to connect all agents and select optimal joint-action for agents; then the learning process (show in Section 3) is based on agents' MDP models and microgrid model to perform the Q-function iteration of each agent; the learning result is each agent's optimal Q-function. Adopting the learning results into practical microgrid market operation (the market model is shown in Section 2), each agent selects current optimal action to participate in the market based on respective current state information and the optimal Q-function.

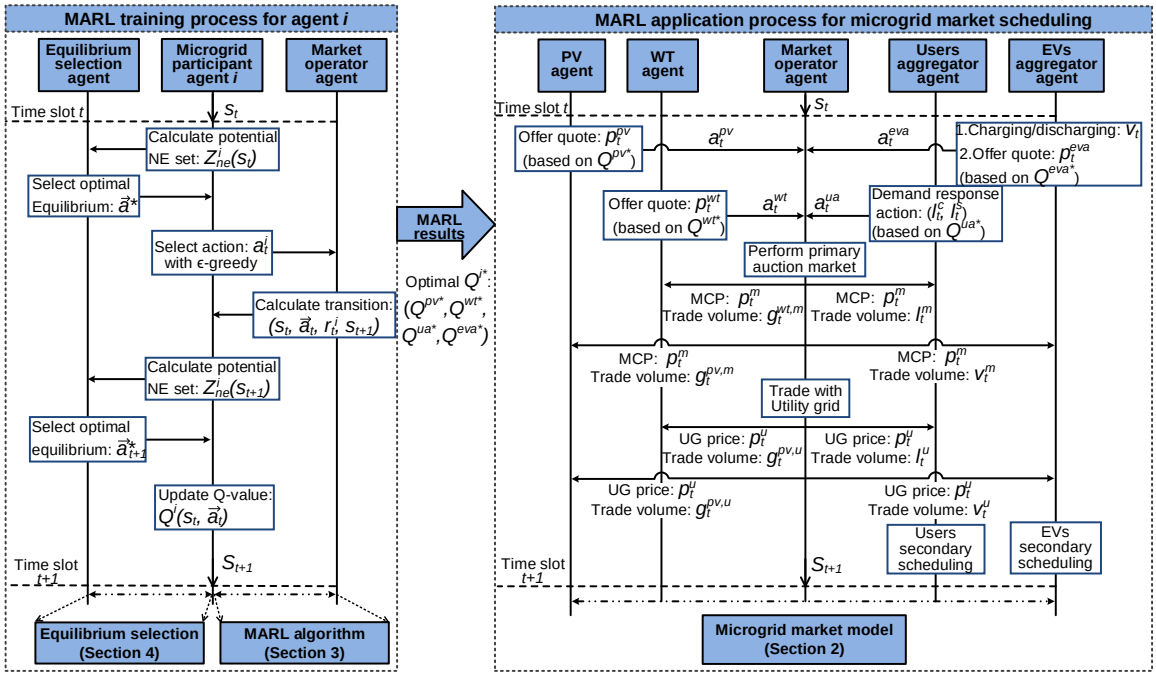

Figure 3. Flowchart of ES-MARL training process and application process for microgrid.

5. Simulation Results and Analysis

In this section, three parts of simulations are conducted to evaluate the proposed MARL algorithm for residential MES. First, the performances of MARL and SARL for MES are compared; then, the effect of proposed ES-MARL is verified and Nash-Q algorithm is used as comparison; finally, the secondary scheduling system of EVs is simulated.

In our microgrid model, an urban residential district is considered. The RGs in the microgrid include one PV and one WT. RGs' daily forecast outputs are extracted from the historical data of a certain area. In microgrid market operation simulations, stochastic models of RGs' generation are used, RGs' actual generation values are generated from probability distributions based on the forecast outputs, the probability distributions of PV and WT are based on [44]; the generation cost functions for PV and WT are $C^{pv}_t = 0.1g^2_t + g_t$ and $C^{wt}_t = 0.1g^2_t + 0.5g_t$. Figure 4a shows the real-time energy purchase price from UG; fiducial forecast values of PV output, WT output; and users' total daily load demand. The number of users is 3; user utility function is $U(d^t_i)$, where the interval of ω is $[1, 4]$, and the value of ω is high or low correspond to load peak period or load trough period, and $\beta = 0.5$.

In our microgrid, there are 10 EVs, EVs' parameters are shown in Table 1. Besides, from [45], the number of arriving EVs or departing EVs at slot t follows normal distribution $\mathcal{N}(n^{arr}_t, 1^2)$ and $\mathcal{N}(n^{dep}_t, 1^2)$, n^{arr}_t and n^{dep}_t are standard values shown as Figure 4b. EVs battery SOC bound is between 0.2 and 0.9; the SOC of arrivals is sampled from $\mathcal{N}(0.5, 0.177^2)$; the travel distance of EV D follow a log-normal distribution $\ln D \sim \mathcal{N}(1.79, 1.09^2)$. All simulations are conducted using Matlab 2018a on the personal computer with Inter Core i7-6700 CPU @3.40GHz. ϵ-greedy strategy is adopted in MARL for action selection strategy, $\epsilon = 1/\ln t$; other learning parameters in RL are shown in Table 2.

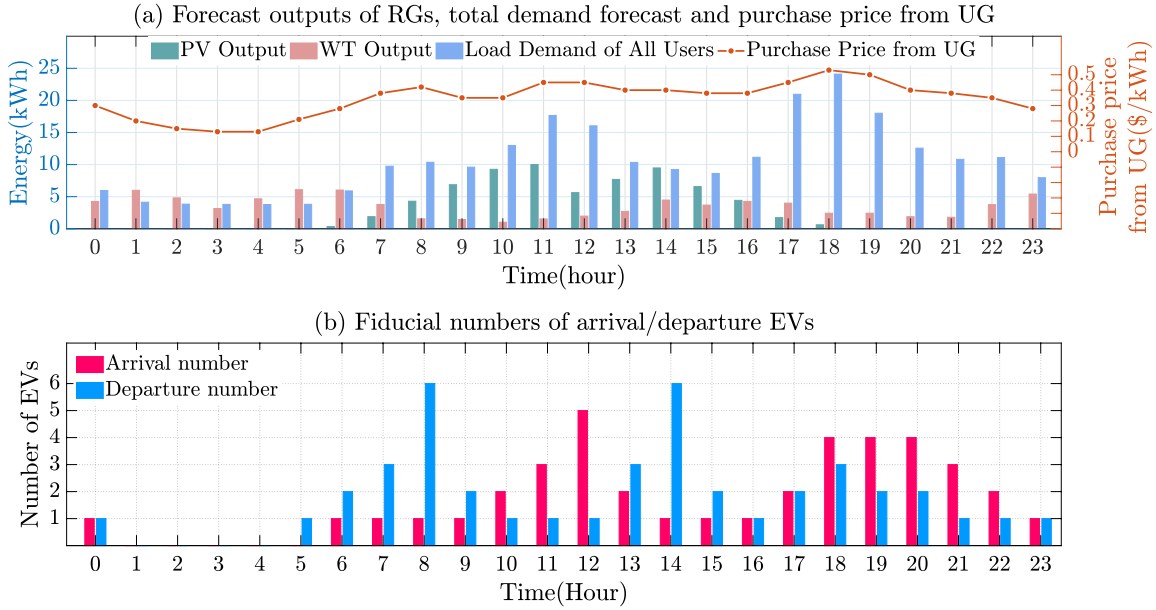

Figure 4. RGs' forecast outputs, total demand forecast, purchase price from UG, fiducial numbers of arrival/departure EVs.

Table 1. EVs' parameters.

Parameter \ EV Type	Nissan Leaf	Buick Velite6	BYD Yuan
Number	4	3	3
Battery capacity	38 kWh	35 kWh	42 kWh
Slow-charge/discharge rating	5 kWh	5.83 kWh	6 kWh
Battery cost	500 $/kWh	500 $/kWh	500 $/kWh

Table 2. Parameters of RL process.

Parameter	PV Agent	WT Agent	UA Agent	EVA Agent
learning rate α	0.7	0.7	0.8	0.8
discount factor γ	0.5	0.5	0.9	0.9

5.1. Performance Comparison of MARL and SARL

5.1.1. Each Agent's Benefit in Different RL Methods

To verify that MARL structure is more applicable than SARL structure in the microgrid market, we simulate the operation of the adopted microgrid model based on the learning results of proposed ES-MARL and various SARL configurations. In SARL, agents can only use public information and their information to learn optimal strategies, and they estimate other agents' privacy information according to experience knowledge. Besides, the agent's learning objective in SARL aims to maximize individual benefit. As contrasts, five SARL configurations are used: in the former four systems, only one kind of agent has learning ability (SARL-PV only, SARL-WT only, SARL-UA only and SARL-EVA only), microgrid market operates with the optimal strategy of learning ability agent, other agents adopt fixed action based on current time slots; in the last SARL, all agents have individual learning ability (SARL-all agents), the market works with selfish optimal strategies from each agent's SARL. We separately evaluate four agent's daily profit in different configurations, the results of stochastic microgrid operation lasting for 30 days are shown in Figures 5–8. In Figures 5 and 6, daily profits of

RGs $Pro^{rg} = \sum_{t=0}^{24} r_t^{rg}$; in Figure 7, daily welfare of UA $W^{ua} = \sum_{t=0}^{24} \sum_{i=1}^{n_u} r_t^{ua}$; in Figure 8, daily total expense of EVA agent is the difference between total electricity purchase cost and total sale income.

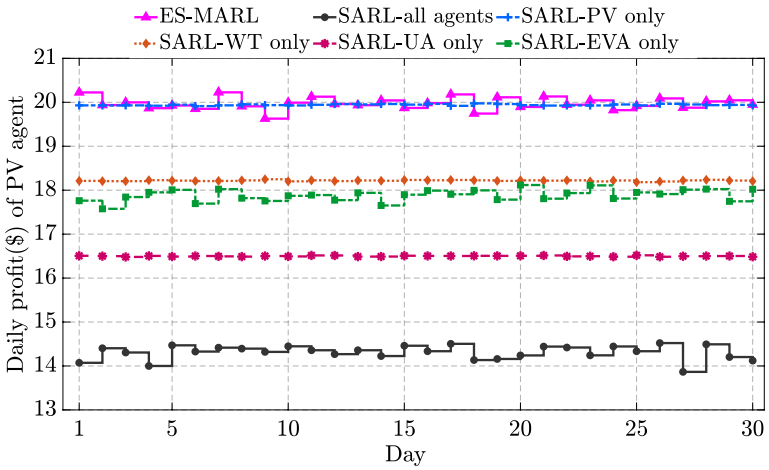

Figure 5. PV agent's daily profit in different RL configurations.

From Figures 5–8, we can see if only one agent has RL ability, the profit of this agent is always highest (or expense is lowest), agent with RL ability can make the optimal decisions based on current market state, but other agents' fixed actions are not optimal for increasing their profit. Moreover, the effects of the proposed ES-MARL method keep in second place for all the four agents. Considering the benefit balance of all agents, it is reasonable that the result of MARL is not as good as selfish SARL for only one agent. However, the global performance of MARL for all agents is optimal. Besides, there are some other notable results. The profits for all agents in SARL-all agents case are the worst, the reason is that in this case, all agents make actions based on their selfish learning results which only care about the self-benefit, this will lead to the imbalance of market and reduce global benefits. We also can find that the curves of SARL-PV only, SARL-WT only and SARL-UA only are more stable (EVA without learning ability), a reasonable explanation is that the randomness of EVs is far more than other members, with RL learning ability, EVA performs actions with the random states of EVs, so the market scheduling results will fluctuate.

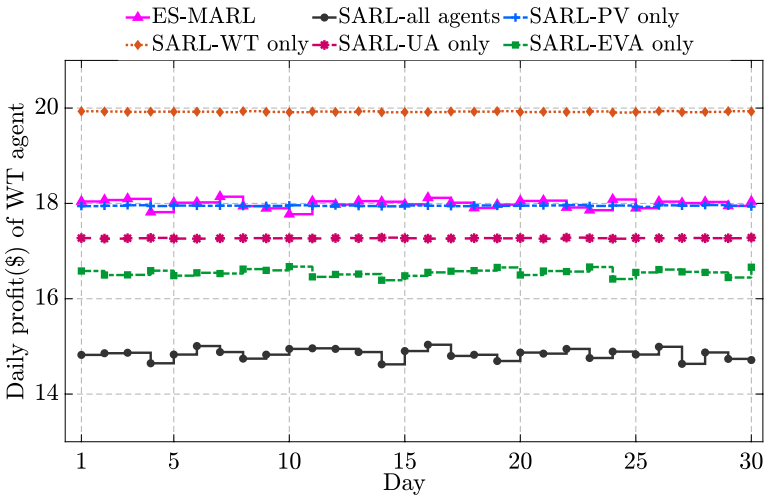

Figure 6. WT agent's daily profit in different RL configurations.

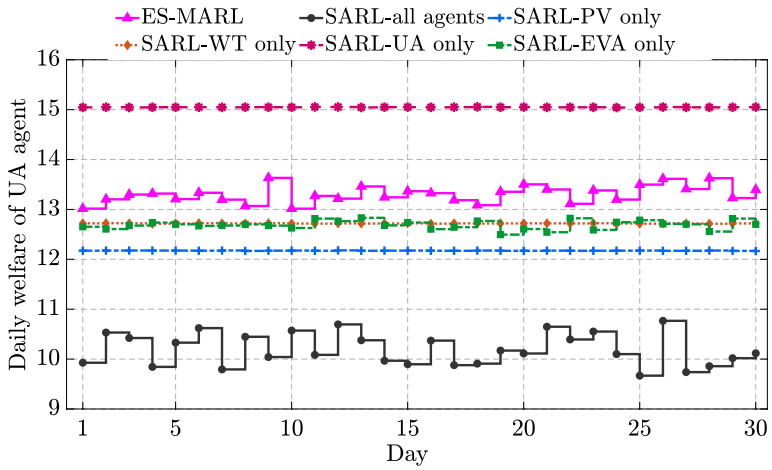

Figure 7. UA agent's daily welfare in different RL configurations.

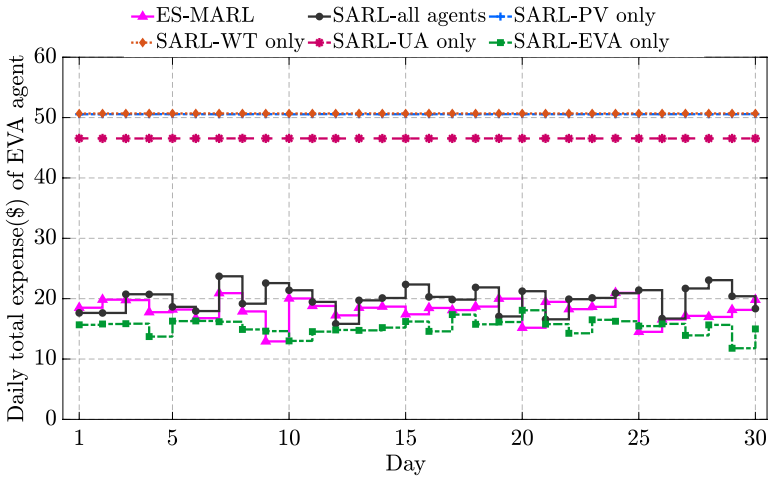

Figure 8. EVA agent's daily expense in different RL configurations.

5.1.2. Overall Performance in Different Microgrid Configurations

For an MES, market fairness and microgrid independence are the two most important indicators. Fairness is to guarantee all participants' benefits achieving equilibrium; microgrid independence aims to realize the supple-demand balance and reduce dependency on UG. Therefore, two global indexes, agents' average profit and daily energy purchase from UG, are introduced to evaluate overall performance. Agents' average profit indicates the overall benefit of microgrid operation; daily energy purchase from UG indicates the dependence level of microgrid on UG. Moreover, to verify the algorithm validity in general cases, two different microgrid configurations are adopted for operation with different RL methods. Microgrid configuration 1: one PV, one WT, 3 users and 10 EVs; microgrid configuration 2: 3 PV, 2 WT, 8 users and 20 EVs. The results are shown in Figures 9–11.

As depicted in Figure 9, agents' average profit is highest with ES-MARL in the two configurations, the value keeps between 30 and 40 for microgrid 1 and 50 to 70 for microgrid 2. This result indicates that ES-MARL produces the best performance to maximize global benefit comparing to other SARL methods. In addition, we can see that the average profits in SARL-PV only, SARL-WT only and SARL-UA only are almost close to zero, the reason is that the demand of EVs' charge is higher than RGs' supply, if EVA has no learning ability to make optimal decisions, the charging cost of EVs is expensive, therefore, the average benefit is offset by EVs' charge expense in these three cases.

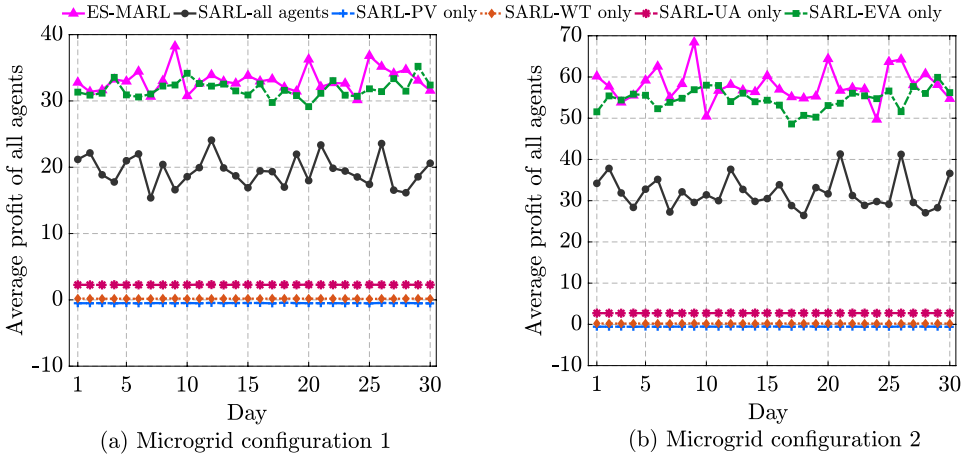

Figure 9. Agents' average profits different RL configurations and different microgrid configurations.

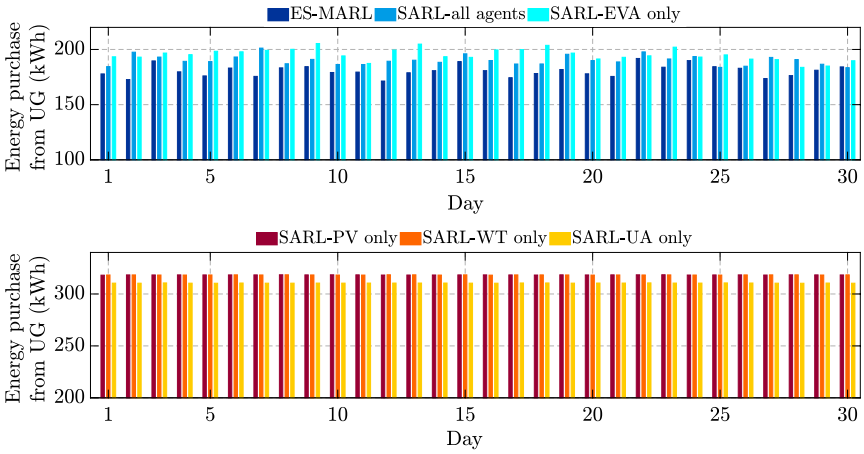

Figure 10. Daily energy purchase from UG in different RL configurations for microgrid configuration 1.

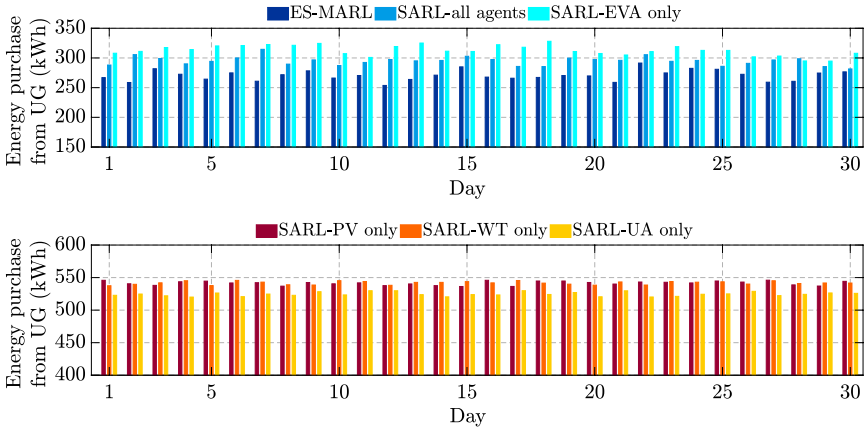

Figure 11. Daily energy purchase from UG in different RL configurations for microgrid configuration 2.

Figures 10 and 11 illustrate the microgrid has the best independence with ES-MARL, the energy purchases from UG in the two configurations are both minimal. The learning process of MARL is based on joint learning, sellers and buyers can reach to appropriate equilibrium to balance the market, agents learn from each other to decrease the supply-demand gap, the microgrid doesn't need to buy more expensive energy from UG. The lower half of Figures 10 and 11 (energy purchase from UG are higher in these cases) also show the importance of EVA agent's learning ability. Combined Figures 10 and 11, the ES-MARL method has the best performance in energy trade with UG; besides, different microgrid models don't affect the final performance of our approach.

5.2. Performance of Improved ES-Based MARL

In this section, several simulations are conducted to evaluate our improved ES-based MARL algorithm. Here, we use Nash-Q MARL algorithm as a comparison, Nash-Q algorithm is the most commonly used MARL. The main difference between Nash-Q and proposed ES-MARL is the equilibrium selection in value function update. We study the algorithm performance from two aspects: performance in the learning process and application effect in residential MES. The RL parameters in two algorithms are set as same as Table 2.

First, four agents' learning performances of the iterative process are shown in Figures 12 and 13. In Figures 12 and 13, the label of the x-axis is episode, which denotes one state transition period from initial state to terminate state, in this paper, one episode is equal to one day (0:00–23:00). The label of the y-axis is Q-value, which is the updated value of Q-function in the current slot. From the results of four agents' Q-values, we can see, the Q-values of ES-MARL is higher than that of Nash-Q throughout the learning process. A bigger Q-value means that the agent's current reward is higher, so ES-MARL can gain a better strategy to increase reward than Nash-Q. Besides, from the curves' trends of four the agents, there is a clear gap in the convergence rates between two algorithms. The state-spaces of the four agents are different (PV's and WT's are smallest, EVA's is biggest), therefore, convergence speeds are accordingly different. For PV and WT, ES-MARL reaches a stable value when $episode = 1.5 \times 10^4$, however, Nash-Q reaches a stable value when $episode = 2 \times 10^4$. For UA and EVA, the convergence episode of ES-MARL is about 2×10^4 and 3×10^4, but when $episode > 3.5 \times 10^4$, the curve of Nash-Q trends to a stable value. The above results show that the convergence speed of ES-MARL is faster.

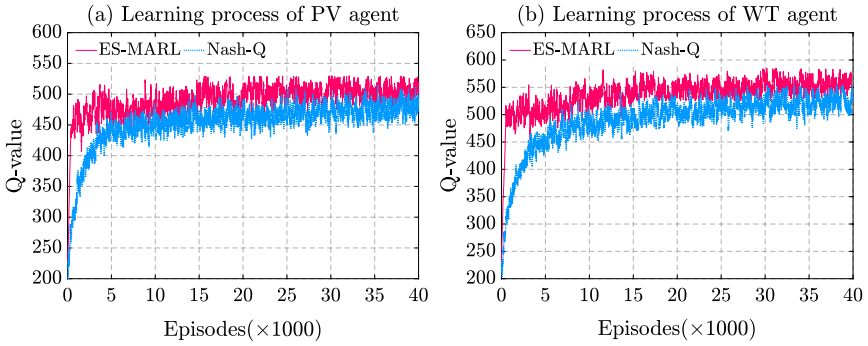

Figure 12. MARL learning process of PV agent and WT agent.

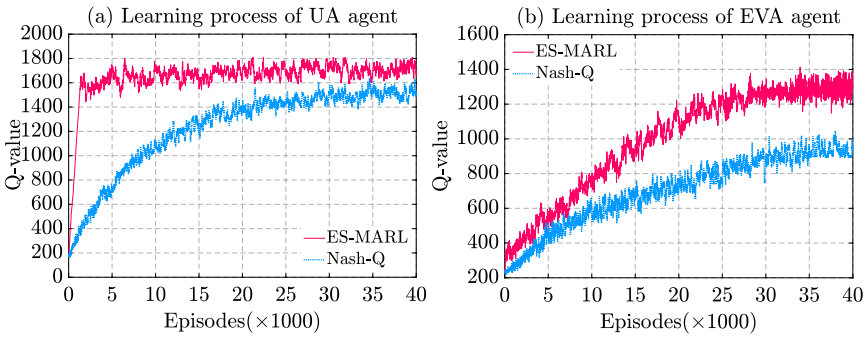

Figure 13. MARL learning process of UA agent and EVA agent.

Then, the results shown in Figures 14–17 illustrate the application performances of the two algorithms when their learning results are adopted in the microgrid market operation. In Figures 14–16, we simulate a one-day microgrid operation with ES-MARL and Nash-Q. The data points of all simulations are calculated by the average value of 100 Monte Carlo experiments. Figure 14 shows the hourly profit of PV and WT. The profits of PV and WT with ES-MARL are superior than with Nash-Q in most hours. The total profits are higher for ES-MARL. Figure 15 depicts the results for UA agent, including two indicators, total energy purchase expense and total welfare (difference between

users' total utility function and total expense), the results show that with ES-MARL, the expense is lower and welfare is higher for UA agent. EVs are both seller and purchaser in the market, the total charge expense and discharge profit of EVA agent are shown in Figure 16. The performance of ES-MARL is better than Nash-Q for less expense and higher profit. To summarize the above analysis, the application performance of ES-MARL is overall better than Nash-Q for all agents. Finally, turn to energy trade between microgrid and UG, consider that energy trade is not the same in the specific hour for different days due to random parameters, we simulate microgrid operation for 30 days as shown in Figure 17. The amount of energy purchase from UG of ES-MARL is less than that of Nash-Q, ES-MARL can reduce microgrid's dependency on UG compared with Nash-Q. Meanwhile, the microgrid will sell back more energy to UG after adopting ES-MARL.

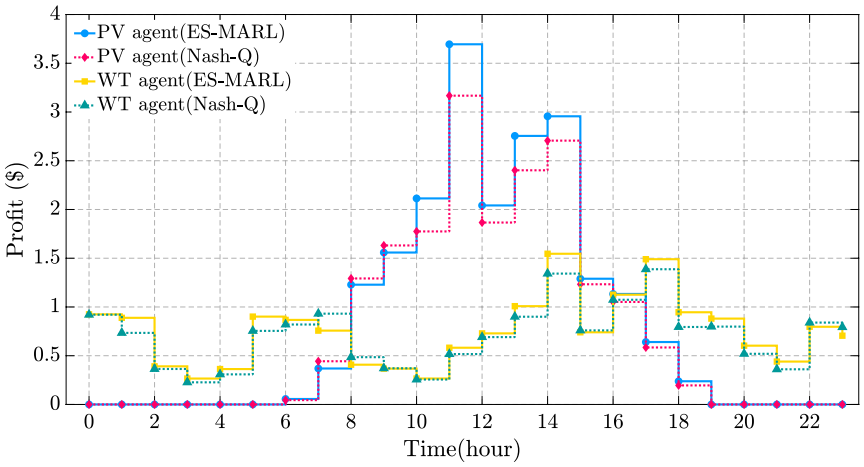

Figure 14. Profit of PV agent and WT agent with ES-MARL and Nash-Q.

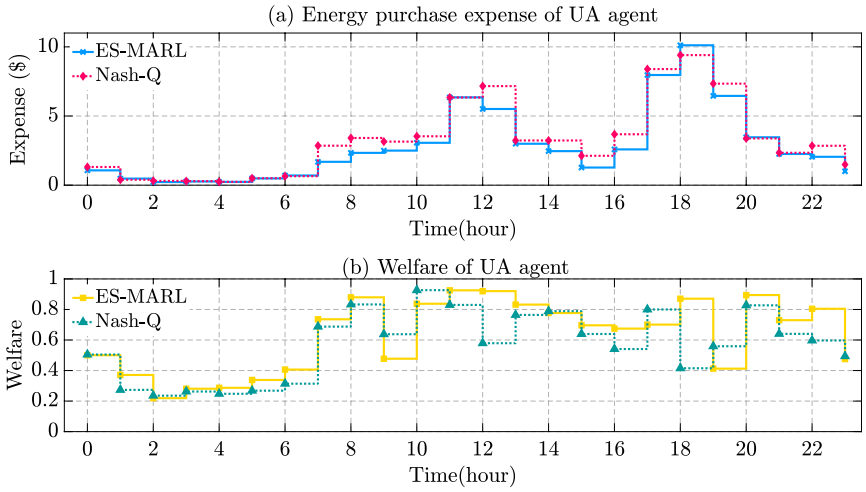

Figure 15. Expense and welfare of UA agent with ES-MARL and Nash-Q.

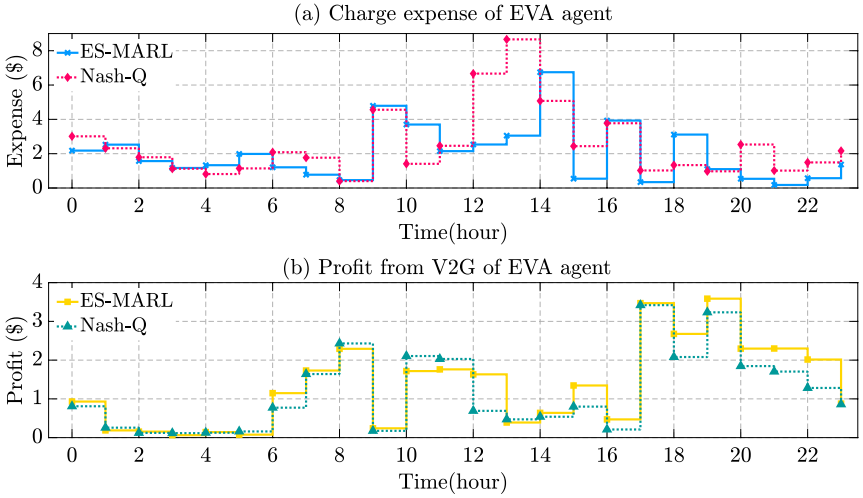

Figure 16. Charge expense and discharge profit of EVA agent with ES-MARL and Nash-Q.

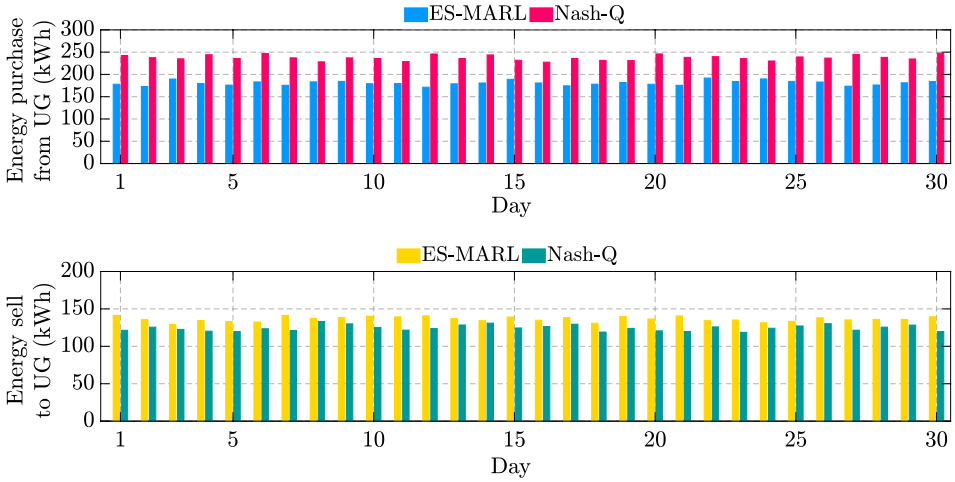

Figure 17. Energy trade with UG (purchase/sell) of ES-MARL and Nash-Q.

5.3. A Simulation Case of EVs Secondary Scheduling System

EVs are important components of a residential microgrid, V2G system plays a crucial role to keep the balance of the microgrid market and enhance microgrid independence. In primary microgrid market, EVA agent represents all EVs to participate in the market, and a secondary scheduling system is set to manager each EV. In this section, we simulate a random secondary scheduling process of EVs for one-day. According to specific EVs' parameters, the primary microgrid market operates to obtain optimal actions for EVA agent in each slot, then EVA arranges each EV's charging/discharging action.

The ten EVs are from different types as shown in Table 1. The set is as follows: Nissan Leaf: EV1-EV4; Buick Velite6: EV5-EV7; BYD Yuan: EV8-EV10. Table 3 shows the EVs' existence state and departure plan. The label "in" denotes that the EV is at home connecting with microgrid in current time; the label "out" means the EV will depart in the next hour (1: "yes"; 0: "no"). The initial time of this case is 0:00. The simulation result of EVs secondary scheduling is shown in Table 4. "Total charge/discharge of EVA" is the optimal decision result from the primary microgrid market. *soc* is EV's SOC state at the beginning of this hour; v_t is the charge/discharge amount, $v_t > 0$ means charge, $v_t < 0$ means discharge, the unit of v_t is kWh. The minimum SOC for departure is 0.8 a.m. and 0.6 p.m. The charge warning limit $soc_{min}^{cha} = 0.3$.

From Tables 3 and 4, the following conclusions can be drawn. First, at each hour, the sum of EV's v_t is equal to the total charge/discharge amount of EVA, which means the secondary scheduling conforms to optimal action from the primary microgrid market. Then, the charging/discharging

sequence is according to the SOC level state, when $soc < 0.3$, the EV is charged immediately. Besides, when EV will depart in current hour, the soc is almost high than the minimum SOC for departure, for example, EV2 will leave at 7:00, so EV2's soc at 7:00 reach to 0.846 (here, the value of v_t is "-" that denotes EV departs at this hour); moreover, EV2 is arranged to deeply charge at 5:00 (two hours ahead). These results can verify the efficiency of EVs' secondary scheduling system.

Table 3. EVs' states parameters.

Hour	EV1		EV2		EV3		EV4		EV5		EV6		EV7		EV8		EV9		EV10	
	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
0:00	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
1:00	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
2:00	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
3:00	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
4:00	1	0	1	0	1	0	1	0	1	0	1	1	1	0	1	0	1	0	1	0
5:00	1	1	1	0	1	0	1	0	1	0	0	-	1	0	1	0	1	0	1	0
6:00	0	-	1	1	1	0	1	1	1	1	0	-	1	0	1	0	1	0	1	0
7:00	0	-	0	-	1	1	0	-	0	-	0	-	1	1	1	1	1	0	1	1
8:00	0	-	0	-	0	-	0	-	0	-	1	0	0	-	0	-	1	1	0	-
9:00	0	-	0	-	0	-	0	-	0	-	1	0	0	-	0	-	0	-	0	-
10:00	1	0	0	-	0	-	0	-	0	-	1	0	0	-	0	-	0	-	0	-
11:00	1	0	1	0	0	-	0	-	0	-	1	0	0	-	1	0	0	-	0	-
12:00	1	1	1	0	1	0	1	0	0	-	1	1	0	-	1	0	0	-	0	-
13:00	0	-	1	1	1	0	1	1	1	1	0	-	0	-	1	1	0	-	0	-
14:00	0	-	0	-	1	1	0	-	0	-	0	-	0	-	0	-	0	-	0	-
15:00	0	-	0	-	0	-	0	-	0	-	0	-	0	-	0	-	1	0	0	-
16:00	0	-	1	0	0	-	0	-	0	-	0	-	0	-	0	-	1	1	0	-
17:00	0	-	1	0	1	0	1	0	0	-	0	-	0	-	0	-	0	-	1	0
18:00	1	1	1	0	1	0	1	0	1	0	0	-	1	0	0	-	0	-	1	0
19:00	0	-	1	0	1	0	1	0	1	0	0	0	1	0	1	0	0	-	1	0
20:00	1	0	1	0	1	0	1	0	1	0	0	-	1	1	1	0	0	-	1	0
21:00	1	1	1	0	1	0	1	0	1	0	1	0	0	-	1	0	0	-	1	0
22:00	0	-	1	0	1	0	1	0	1	0	1	0	0	-	1	0	0	-	1	0
23:00	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	-	1	0

Table 4. A case of EVs secondary scheduling system.

Hour	Total Charge/ Discharge of EVA	EV1		EV2		EV3		EV4		EV5		EV6		EV7		EV8		EV9		EV10	
		soc(%)	v_t	soc(%)	v_t	soc(%)	v_t	soc(%)	v_t	soc(%)	v_t	soc(%)	v_t	soc(%)	v_t	soc	v_t	soc(%)	v_t	soc(%)	v_t
0:00	5	22	5	27	5	73	−3.41	54	−1.51	23	5	82	−3.41	35	1.74	44	0	61	−3.41	45	0
1:00	10	35.2	3	40.2	0.5	64	0.5	50	0.5	39.7	3	72.3	0.5	40	0.5	44	0.5	52.9	0.5	45	0.5
2:00	20	43.2	2	41.5	2	65.3	2	51.3	2	48.3	2	74.7	2	41.4	2	45.2	2	54.1	2	46.2	2
3:00	20	48.5	2	46.8	2	70.6	2	56.6	2	54	2	80.4	2	47.1	2	50	2	58.9	2	51	2
4:00	25	53.8	5	52.1	2.33	75.9	2.33	61.9	2.33	59.7	2.33	86.1	2.33	52.8	2.33	54.8	2.33	63.7	2.33	55.8	2.33
5:00	15	67	5	58.2	5	82	0	68	5	66.4	5	90	-	59.5	0	60.3	0	69.2	−4.45	61.3	−0.55
6:00	20	80.1	-	71.4	5	82	0	81.2	0	80.7	0	-	-	59.5	5	60.3	5	58.6	0	60	5
7:00	15	-	-	84.6	-	82	0	81.2	-	80.7	-	-	-	71.8	3.3	72.2	3.3	58.6	5	71.9	3.4
8:00	10	-	-	-	-	-	-	-	-	-	-	61	5	83.2	-	80	-	70.5	5	80	-
9:00	5	-	-	-	-	-	-	-	-	-	-	75.3	5	-	-	-	-	82.4	-	-	-
10:00	−5	63	−2.5	-	-	-	-	-	-	-	-	89.6	−2.5	-	-	-	-	-	-	-	-
11:00	−10	56.4	3.9	64	−5	-	-	-	-	-	-	82.5	−5	-	-	59.3	−3.9	-	-	-	-
12:00	15	66.7	5	50.8	1.96	53	0	42	1.96	-	-	68.2	4.13	-	-	50	1.96	-	-	-	-
13:00	20	79.9	-	56	5	53	5	47.6	5	-	-	80	-	-	-	54.7	5	-	-	-	-
14:00	5	-	-	69.2	-	66.2	5	61.9	-	-	-	-	-	-	-	66.7	-	-	-	-	-
15:00	0	-	-	-	-	79.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
16:00	−5	-	-	53	−5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
17:00	−10	-	-	39.8	0	71.2	−5	47.5	−2.5	-	-	-	-	-	-	-	-	-	-	52.1	-2.5
18:00	−15	37.5	0	39.8	0	58	−5	40.9	−4.14	28	4.14	-	-	61	−5	-	-	-	-	46.1	−5
19:00	−5	37.5	−1.67	39.8	−1.67	44.8	−1.67	30	0	39.8	−1.67	-	-	46.7	5	42.1	−1.67	-	-	34.2	−1.67
20:00	−5	33.1	−0.83	35.4	−0.83	40.4	−0.83	30	0	35	−0.83	-	-	61	-	38.1	−0.83	-	-	30.2	−0.83
21:00	−5	30.9	0	33.2	0	38.2	0	30	0	32.6	0	71	−5	-	-	36.1	0	-	-	28.2	0
22:00	10	30.9	1.43	33.2	1.43	38.2	1.43	30	1.43	32.6	1.43	56.7	0	-	-	36.1	1.43	-	-	28.2	1.43
23:00	10	34.6	1.43	37	1.43	42	1.43	34.1	1.43	36.7	1.43	56.7	0	54	0	39.5	1.43	-	-	31.6	1.43

6. Conclusions

In this paper, we concentrate on the energy scheduling of residential microgrid. The integrated residential microgrid system including RGs, power users and EVs V2G is constructed on the multi-agent structure and an auction-based microgrid market mechanism is built to adapt microgrid participants' demands for distributed management and independent decision.

In order to generate the optimal market strategy for each participant and to guarantee the balance of all participants' benefit and microgrid supply-demand, we introduce a model-free MARL approach for each agent. Through MARL, agents can consider both farsighted self-interest and the actions of other agents to make decision in a dynamic and stochastic market environment. Moreover, we present a novel ES-MARL algorithm to improve the privacy security, fairness, and efficiency of MARL. There are two cardinal mechanisms in ES-MARL, one is private negotiation between ESA and each agent, which can protect private information and reduce computational complexity; another is the maximum average reward method to select a global optimal equilibrium solution.

Three parts of simulations have been carried out: (1) the comparison results between MARL and SARL verify that MARL is more appropriate for distributed microgrid scheduling to ensure agents individual benefits and overall operation objective; (2) the simulations with proposed ES-MARL and classic Nash-Q MARL are conducted, the results show that our proposed approach can achieve better performance of learning process and microgrid application; (3) a case study about 10 EVs charging/discharging scheduling demonstrates the effectiveness of secondary EVs scheduling system.

In conclusion, this work adopts an improved MARL approach for residential microgrid market scheduling. The learning results can enable agents to autonomously select strategy for promoting benefit; meanwhile, the microgrid system can coordinate all participant's demands and achieve a high autonomy under the equilibrium-based learning process.

Author Contributions: Conceptualization, X.F., J.W. and Y.H.; Methodology, X.F., Y.H. and Q.Z.; Software, X.F., Z.C. and G.S.; Validation, Y.H. and Q.Z.; Writing—Original Draft Preparation, X.F., G.S. and Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (2016YFB0901900); Natural Science Foundation of Hebei Province (F2017501107); Open Research Fund from the State Key Laboratory of Rolling and Automation, Northeastern University (2017RALKFKT003); Fundamental Research Funds for the Central Universities (N182303037); Foundation of Northeastern University at Qinhuangdao (XNB201803).

Acknowledgments: The authors gratefully acknowledge the National Key Research and Development Program of China.

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

References

1. Sattarpour, T.; Golshannavaz, S.; Nazarpour, D.; Siano, P. A multi-stage linearized interactive operation model of smart distribution grid with residential microgrids. *Int. J. Electr. Power Energy Syst.* **2019**, *108*, 456–471. [CrossRef]
2. Pascual, J.; Barricarte, J.; Sanchis, P.; Marroyo, L. Energy management strategy for a renewable-based residential microgrid with generation and demand forecasting. *Appl. Energy* **2015**, *158*, 12–25. [CrossRef]
3. Zhang, X.; Bao, J.; Wang, R.; Zheng, C.; Skyllas-Kazacos, M. Dissipativity based distributed economic model predictive control for residential microgrids with renewable energy generation and battery energy storage. *Renew. Energy* **2017**, *100*, 18–34. [CrossRef]
4. Wang, H.; Zhang, X.; Ouyang, M. Energy consumption of electric vehicles based on real-world driving patterns: A case study of Beijing. *Appl. Energy* **2015**, *157*, 710–719. [CrossRef]
5. Zhou, G.; Ou, X.; Zhang, X. Development of electric vehicles use in China: A study from the perspective of life-cycle energy consumption and greenhouse gas emissions. *Appl. Energy* **2013**, *59*, 875–884. [CrossRef]
6. Rodrigues, Y.R.; de Souza, A.Z.; Ribeiro, P. An inclusive methodology for Plug-in electrical vehicle operation with G2V and V2G in smart microgrid environments. *Int. J. Electr. Power Energy Syst.* **2018**, *102*, 312–323. [CrossRef]

7. Wang, K.; Gu, L.; He, X.; Guo, S.; Sun, Y.; Vinel, A.; Shen, J. Distributed Energy Management for Vehicle-to-Grid Networks. *IEEE Netw.* **2017**, *31*, 22–28. [CrossRef]
8. Thomas, D.; Deblecker, O.; Ivatloo, B.M.; Ioakimidis, C. Optimal operation of an energy management system for a grid-connected smart building considering photovoltaics' uncertainty and stochastic electric vehicles' driving schedule. *Appl. Energy* **2018**, *210*, 1188–1206. [CrossRef]
9. 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. [CrossRef]
10. Foruzan, E.; Soh, L.; Asgarpoor, S. Reinforcement Learning Approach for Optimal Distributed Energy Management in a Microgrid. *IEEE Trans. Power Syst.* **2018**, *33*, 5749–5758. [CrossRef]
11. Hu, Y.; Gao, Y.; An, B. Multiagent Reinforcement Learning With Unshared Value Functions. *IEEE Trans. Cybern.* **2015**, *45*, 647–662. [CrossRef] [PubMed]
12. Wang, H.; Huang, T.; Liao, X.; Abu-Rub, H.; Chen, G. Reinforcement Learning for Constrained Energy Trading Games With Incomplete Information. *IEEE Trans. Cybern.* **2017**, *47*, 3404–3416. [CrossRef] [PubMed]
13. Zhou, L.; Yang, P.; Chen, C.; Gao, Y. Multiagent Reinforcement Learning With Sparse Interactions by Negotiation and Knowledge Transfer. *IEEE Trans. Cybern.* **2017**, *47*, 1238–1250. [CrossRef] [PubMed]
14. Vasirani, M.; Kota, R.; Cavalcante, R.L.G.; Ossowski, S.; Jennings, N.R. An Agent-Based Approach to Virtual Power Plants of Wind Power Generators and Electric Vehicles. *IEEE Trans. Smart Grid* **2013**, *4*, 1314–1322. [CrossRef]
15. Shamsi, P.; Xie, H.; Longe, A.; Joo, J. Economic Dispatch for an Agent-Based Community Microgrid. *IEEE Trans. Smart Grid* **2016**, *7*, 2317–2324. [CrossRef]
16. Li, C.; Ahn, C.; Peng, H.; Sun, J. Synergistic control of plug-in vehicle charging and wind power scheduling. *IEEE Trans. Power Syst.* **2013**, *28*, 1113–1121. [CrossRef]
17. Karfopoulos, E.L.; Hatziargyriou, N.D. Distributed Coordination of Electric Vehicles Providing V2G Services. *IEEE Trans. Power Syst.* **2016**, *31*, 329–338. [CrossRef]
18. Marzband, M.; Javadi, M.; Pourmousavi, S.A.; Lightbody, G. An advanced retail electricity market for active distribution systems and home microgrid interoperability based on game theory. *Electr. Power Syst. Res.* **2018**, *157*, 187–199. [CrossRef]
19. Wang, C.; Liu, Y.; Li, X.; Guo, L.; Qiao, L.; Lu, H. Energy management system for stand-alone diesel-wind-biomass microgrid with energy storage system. *Energy* **2016**, *97*, 90–104. [CrossRef]
20. Marzband, M.; Alavi, H.; Ghazimirsaeid, S.S.; Uppal, H.; Fernando, T. Optimal energy management system based on stochastic approach for a home Microgrid with integrated responsive load demand and energy storage. *Sustain. Cities Soc.* **2017**, *28*, 256 – 264. [CrossRef]
21. Kim, B.; Zhang, Y.; van der Schaar, M.; Lee, J. Dynamic Pricing and Energy Consumption Scheduling With Reinforcement Learning. *IEEE Trans. Smart Grid* **2016**, *7*, 2187–2198. [CrossRef]
22. Ruelens, F.; Claessens, B.J.; Vandael, S.; Schutter, B.D.; Babuska, R.; Belmans, R. Residential Demand Response of Thermostatically Controlled Loads Using Batch Reinforcement Learning. *IEEE Trans. Smart Grid* **2017**, *8*, 2149–2159. [CrossRef]
23. 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. [CrossRef]
24. Silva, M.; Souza, S.; Souza, M.; Bazzan, A. A reinforcement learning-based multi-agent framework applied for solving routing and scheduling problems. *Expert Syst. Appl.* **2019**, *131*, 148–171. [CrossRef]
25. Kazmi, H.; Suykens, J.; Balint, A.; Driesen, J. Multi-agent reinforcement learning for modeling and control of thermostatically controlled loads. *Appl. Energy* **2019**, *238*, 1022–1035. [CrossRef]
26. Littman, M.L. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994, Proceedings of the Eleventh International Conference, Rutgers University, New Brunswick, NJ, 10–13 July 1994*; Morgan Kaufmann: Amsterdam, The Netherlands, 1994; pp. 157–163.
27. Hu, J.; Wellman, M.P. Nash Q-Learning for General-Sum Stochastic Games. *J. Mach. Learn. Res.* **2003**, *4*, 1039–1069.
28. Littman, M.L. Friend-or-foe Q-learning in general-sum games. In Proceedings of the Eighteenth International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001; Morgan Kaufmann: Amsterdam, The Netherlands, 2001.

29. Greenwald, A.; Hall, K. Correlated Q-Learning. In Proceedings of the Twentieth International Conference on International Conference on Machine Learning, Washington, DC, USA, 21–24 August 2003; AAAI Press: New Orleans, LA, USA, 2003; pp. 242–249 .
30. Bowling, M.; Veloso, M. Multiagent learning using a variable learning rate. *Artif. Intell.* **2002**, *136*, 215–250. [CrossRef]
31. Rahman, M.; Oo, A. Distributed multi-agent based coordinated power management and control strategy for microgrids with distributed energy resources. *Energy Convers. Manag.* **2017**, *139*, 20–32. [CrossRef]
32. Kofinas, P.; Dounis, A.; Vouros, G. Fuzzy Q-Learning for multi-agent decentralized energy management in microgrids. *Appl. Energy* **2018**, *219*, 53–67. [CrossRef]
33. Nunna, H.S.V.S.K.; Battula, S.; Doolla, S.; Srinivasan, D. Energy Management in Smart Distribution Systems With Vehicle-to-Grid Integrated Microgrids. *IEEE Trans. Smart Grid* **2018**, *9*, 4004–4016. [CrossRef]
34. BU-205: Types of Lithium-ion. Available online: http://www.batteryuniversity.com/learn/article/types_of_lithium_ion (accessed on 18 September 2019).
35. Ortega-Vazquez, A.M. Optimal scheduling of electric vehicle charging and vehicle-to-grid services at household level including battery degradation and price uncertainty. *IET Gener. Transm. Distrib.* **2014**, *8*, 1007–1016. [CrossRef]
36. Igualada, L.; Corchero, C.; Cruz-Zambrano, M.; Heredia, F. Optimal Energy Management for a Residential Microgrid Including a Vehicle-to-Grid System. *IEEE Trans. Smart Grid* **2014**, *5*, 2163–2172. [CrossRef]
37. Yona, A.; Senjyu, T.; Funabashi, T. Application of recurrent neural network to short-term-ahead generating power forecasting for photovoltaic system. *IEEE Power Eng. Soc. Gen. Meet.* **2007**, *86*, 3659–3664.
38. Borowy, B.; Salameh, Z. Methodology for Optimally Sizing the Combination of a Battery Bank and PV Array in a Wind/PV Hybrid System. *IEEE Trans. Energy Convers.* **1996**, *11*, 367–375. [CrossRef]
39. Li, T.; Dong, M. Residential Energy Storage Management With Bidirectional Energy Control. *IEEE Trans. Smart Grid* **2019**, *10*, 3596–3611. [CrossRef]
40. Cintuglu, M.H.; Martin, H.; Mohammed, O.A. Real-Time Implementation of Multiagent-Based Game Theory Reverse Auction Model for Microgrid Market Operation. *IEEE Trans. Smart Grid* **2015**, *6*, 1064–1072. [CrossRef]
41. Hu, J.; Wellman, M.P. Multiagent Reinforcement Learning: Theoretical Framework and an Algorithm. In Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98), Madison, WI, USA, 24–27 July 1998; Morgan Kaufmann: Amsterdam, The Netherlands, 1998; pp. 242–250.
42. Buoniu, L.; Babuka, R.; Schutter, B.D., Multi-agent Reinforcement Learning: An Overview. In *Innovations in Multi-Agent Systems and Applications–1*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 183–221.
43. Ko, H.; Pack, S.; Leung, V.C.M. Mobility-Aware Vehicle-to-Grid Control Algorithm in Microgrids. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 2165–2174. [CrossRef]
44. Atwa, Y.M.; El-Saadany, E.F.; Salama, M.M.A.; Seethapathy, R. Optimal Renewable Resources Mix for Distribution System Energy Loss Minimization. *IEEE Trans. Power Syst.* **2010**, *25*, 360–370. [CrossRef]
45. Yao, L.; Lim, W.H.; Tsai, T.S. A Real-Time Charging Scheme for Demand Response in Electric Vehicle Parking Station. *IEEE Trans. Smart Grid* **2017**, *8*, 52–62. [CrossRef]