IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 5, 2020, accepted June 8, 2020, date of publication June 16, 2020, date of current version June 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3002693

Probability-Based Energy Reinforced Management of Electric Vehicle Aggregation in the Electrical Grid Frequency Regulation

CHAOYU DONG[1,2,4], JIANWEN SUN[3], FENG WU[3], (Member, IEEE), AND HONGJIE JIA[1,2], (Senior Member, IEEE)

[1]Key Laboratory of Smart Grid, Ministry of Education, Tianjin University, Tianjin 300072, China
[2]Energy Research Institute, Nanyang Technological University, Singapore 637141
[3]School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798
[4]Energy Research Institute, Nanyang Technological University, Singapore 637141

Corresponding author: Jianwen Sun (kevensun@ntu.edu.sg)

This work was supported in part by the National Natural Science Foundation of China (U1766210, 51377117, 51625702).

ABSTRACT The model uncertainties and the heterogeneous energy states burden the effective aggregation of electric vehicles (EVs), especially coupling with the real-time frequency dynamic of the electrical grid. Integrating the advantages of deep learning and reinforcement learning, deep reinforcement learning shows its potential to relieve this challenge, where an intelligent agent fully considers the individual state of charge (SOC) difference of EV and the grid state to optimize the aggregation performance. However, existing policies of deep reinforcement learning usually provide deterministic and certain actions, and it is difficult to deal with the increasing uncertainties and randomness in modern electrical systems. In this paper, a probability-based management strategy is proposed with continuous action space based on the deep reinforcement learning, which provides fine-grained energy management and addresses the time-varying dynamics from EVs and electrical grid simultaneously. Moreover, an optimization based on the proximal policy is further introduced to clip the policy upgradation speed to enhance the training stability. The effectiveness of proposed energy management structure and policy optimization strategy are verified on various scenarios and uncertainties, which demonstrates advantageous performance in the SOC management and frequency maintenance. Besides the performance merits, the training procedure is also presented revealing the evolution reason for the proposed approach.

INDEX TERMS State of charge (SOC), deep reinforcement learning, hybrid framework, electric vehicle aggregator, multiple-input and multiple-output.

NOMENCLATURE

α_π	Learning stepsizes of policy network
α_v	Learning stepsizes of value function network
β	Bias factor
Δf	Frequency variation of electrical grid
ΔSOC_i	SOC variation of Aggregator i
ΔP_g	Turbine power variation
ΔP_L	Load imbalance
ΔP_m	Reheat power variation
ΔX_g	Valve position
γ	Discount factor
$\hat{A}$	Advantage estimator
$\mathbf{S}_{grid}$	Observed grid vector
$\mathbf{S}_{in}$	Observation vector
$\mathbf{S}_i^{EV}$	Observed vector of Aggregator i
$\mathbf{U}_{in}$	Policy vector
$\mathcal{H}$	Trajectory storage
μ	Mean value
π	Management policy
π^*	Optimized policy network
π_θ	Current policy
π_{old}	Old policy
θ	Network parameter of policy network
σ	Variance value
ψ	Network parameter of policy network

The associate editor coordinating the review of this manuscript and approving it for publication was Giambattista Gruosso.

ε	Clipping degree
a_t	Action at timestep t
ACE	Area Control Error
D	Grid damping coefficient
DQN	Deep Q-Network
DRL	Deep reinforcement learning
EV	Electric vehicle
F_r	Fraction of total turbine power
J	Expected return
K_I	Integral coefficient of frequency control
K_P	Proportional coefficient of frequency control
K_i^{EV}	Charging/discharging efficiency of aggregator i
L	Episode number
M	System inertia
MDP	Markov decision process
N	Episode step
P_i^{EV}	Power output of Aggregator i
PPO	Proximal policy optimization
R	Droop coefficient
R_t	Cumulative reward at timestep t
r_t	Reward at timestep t
s_t	Observation at timestep t
SOC	State of charge
$SOC_i^{initial}$	SOC initial value of Aggregator i
T_c	Turbine constant
T_r	Reheat constant
T_i^{EV}	Battery time constant of Aggregator i
u_grid	Grid operator command
u_i^{EV}	Power command for Aggregator i
V	Value function
$V2G$	Vehicle-to-grid
w	Likelihood radio

I. INTRODUCTION

Growing electric vehicles (EVs) and their uncertainties are reconstructing the conventional electrical system. Due to the consumption of fossil fuel and the emission of carbon dioxide, vehicle electrification is extending in many countries. Less than a year from the factory announcement, Tesla's Shanghai plant delivered its first electric vehicle on 30 December 2019 [1]. The production rate of this Gigafactory is expected to reach 500,000 per year [2]. According to the CNN Business report, the affordable version of Model 3 and Model Y will be made in this Chinese factory. [3]. Along with Tesla, local brands such as NIO and BYD are also thriving. From the official report [4] of NIO, more than 30,000 EVs have been delivered since 2018. The NIO ES6 has been consecutively occupying first place among the premium midsize sport utility vehicle (SUV). With the batteries manufactured by itself, BYD covers the global EV business from passenger cars, buses, utility vans, and trucks [5]. With the goal of vehicles-consumed fossil fuel reduction, a large number of EVs are irreversibly penetrating the electrical grid operation.

As the interface between the EVs and the electrical grid, various charging devices are installed worldwide. With the wall connector from Tesla [6], the maximum power can reach 11.5kW at a home charging condition. In the meantime, its supercharger stations have been allocated in Asia, North America, and Europe. Empowered by the new architecture, the V3 supercharging of Tesla can support the 250kW peak rate [7], which significantly releases the range constraint. The newly released charging station of NIO can also reach the 105kW charging rate, requiring charging time less than 30 minutes from 20% to 80% state [8]. For the BYD company, it provides several options, such as 3.3kW, 7.0kW, and 40kW. The maximum charging speed can get to 80kW [9]. Considering the striking power and sale promotion, millions of electric vehicles will form a huge energy system interacting with the electrical grid.

Through the widespread electricity charger, the coalition of the EV system and power system is formulated, which fosters various research topics from the electronic power implementation, dispatch strategy design, and market optimization.

As the spotlight, the vehicle-to-grid (V2G) attracts numerous attention discussing the possibility of grid support with EVs [10]–[13]. From the network stability and individual preferences, the report in [10] proposed industrial and market incentives for V2G development. For an EV fleet, [11] designed a distributed coordination in the regulation services, which satisfied the daily energy demand and day-ahead aggregator schedule. Through the conditional value-at-risk constraints, the bidding strategy was developed in [12] to formulate a decision-making tool in day-ahead and real-time markets. To sufficiently utilize the EV capacity and establish the regulation service, the non-cooperative and cooperative games were developed to stimulate the EV interaction and compress the grid fluctuation [13]. Those reports indicate the financial potentials for the implementation of V2G.

In the real-time dispatch aspect, different model-based strategies have been studied to boost regulation performance. Through the fixed structure mixed H_2/H_∞ control, the robustness of the V2G system is enhanced facing the time-varying delays and uncertainties [14].

With the help of fractional order controller, the faster setting time and overshoot reduction are achieved in the frequency regulation of [15]. A fuzzy controller was carried out by [16] to reduce the frequency deviation considering the state of charge (SOC) difference, but the aimed SOC is pre-designed around a certain value. If the SOC goal is different, the original fuzzy rule of the 5*7 table should be re-designed, changing fuzzy regions. Besides, relying on the system model, these algorithms have to re-optimize all the control parameters. Due to the variation of EV devices as well as the time-varying uncertainties of their SOCs, it is necessary to design an energy management strategy free from the complicated model derivation.

In the past years, the rapid development has been witnessed in the field of deep reinforcement learning (DRL) technology, which combines the perception capability of

deep learning and the intelligent decision-making ability of reinforcement learning. Advanced DRL algorithms were developed to achieve remarkable performance on various artificial intelligence takes, e.g., Go [17], robotics [18], video games [19], [20]. Unlike traditional reinforcement learning, the DRL algorithms use powerful deep neuron networks to approximate their value function (such as Q-table), enabling automatic high-dimensional feature extraction and end-to-end learning. Recently, the advantages of DRL were recognized by the community and some attempts were made to leverage DRL in various applications for electrical grid, including operational control [21]–[24], electricity market [25], [26], demand response [27] and energy management [28]. Although these applications presented advantageous results in their respective fields, several challenges were encountered. Some early works [21], [22], [25] adopted deep Q-network as the agent's policy, which could only deal with control problems with discrete action spaces. These applications were not practical for physical regulation tasks in power systems, requiring continuous action spaces. Several works [23], [24], [26]–[28] addressed the continuous control problem by using the deep policy gradient method. However, the policies of these applications directly output deterministic action ignoring the uncertainties and randomness in the electrical grid. The random fluctuations in electrical grids and the incomplete modeling of the control system burden the application of deterministic method. On account of the continuous action space and complicated environment disturbances, it is more practical to use a probabilistic control policy rather than a deterministic one in the integration management of electric vehicles and electrical grid.

To mitigate the uncertainties and randomness in the EV-grid energy management, a probability-based energy management structure is developed for the EV aggregation. The proposed schemes regulates the EV charging/discharging power according to the energy difference in the continuous action space, enabling more flexible and fine-grained control for the system dispatch. Our algorithm framework is integrated with state-of-art deep reinforcement learning algorithms, which effectively elevates the training efficiency and boost the management performance. The main contributions of this work are summarized below.

1. **Energy management structure**. With the detailed derivation of EV and electrical grid dynamics, a general management structure is established for the aggregation of heterogeneous EVs. In the proposed structure, deep reinforcement learning is deployed, coordinating with the traditional automatic generation control to meet the multiple-inputs and multiple-outputs requirement of the electrical grid-electric vehicle system. Besides, the advantage of the deep neural network is sufficiently utilized to observe the high-dimensional grid and EV dynamics. The established structure can provide a unified strategy for the tests of various optimization approaches through the flexible adjustment of the observation and the objective function.

2. **Probabilistic continuous management.** To support continuous control based on reinforcement learning, a probabilistic model with continuous action space is developed to represent the intelligent dispatch center. Instead of outputting the control signals directly, the proposed policy model solves the control problem by approximating the optimal probability distribution of the control signals. The proposed approach can achieve better performance when dealing with the randomness of the electrical grid.

3. **Improved gradient-based policy optimization**. To improve the optimization stability of the training process, an optimization algorithm based on a clipped surrogate objective is employed for energy management optimization. Through the probability ratio between old and new policies, an upgrading constraint is set to guarantee the policy evolution within a specific interval. The proposed algorithm shows stable training performance during the learning process.

The reminder of this paper is organized as follows. Section II provides a detailed literature review regarding deep reinforcement learning and its application in power systems. Section III reviews the dynamics of electric vehicles and the electrical grid, respectively. On account of the modelling difficulty and the time-varying feature, a probabilistic energy reinforced management structure is then proposed in response to the EV dispatch challenge. For the implementation of the management structure, the Markov decision process is introduced in Section IV, which bridges the EV aggregation problem and the deep reinforcement learning. The training algorithm is then detailed to achieve the self-regulation assessment and the policy evolution. The demonstration of energy management performance is shown and analyzed in Section V with a typical electrical grid-electric vehicle system, which reveals the advantageous properties and the improvement reasons. The conclusions are provided in Section VI to summarize this work and results.

II. RELATED WORKS

A. DEEP REINFORCEMENT LEARNING

Reinforcement learning is research filed on machine learning, where the goal is to train an agent to maximize its reward by interacting with the environment. Recently, deep learning has started to play an essential role in reinforcement learning. Researchers used deep neuron networks as ideal function approximators for reinforcement learning algorithms from high dimensional signals for building informed action determination process. DRL is considered a promising artificial intelligence method that is close to human thinking. Many advanced DRL algorithms have been developed in recent years. As an early work, Deep Q-network (DQN) [29] is introduced by using a multi-layer neuron network as the approximation of Q function. To support continuous control, paper [30] proposed the deep deterministic policy gradient (DDPG) algorithm in continuous action spaces by extending DQN to policy gradient method.

TABLE 1. The comparison of related studies.

Works	Application Field	Algorithm	Action Space	Continuous Control	Policy Type	Probabilistic Policy
Huang *et al.* [21]	Operational Control	DQN	Discrete	×	Deterministic	×
Chen and Su [25]	Electricity Market	DQN	Discrete	×	Deterministic	×
Yang *et al.* [22]	Operational Control	DQN	Discrete	×	Deterministic	×
Yan and Xu [23]	Operational Control	DDPG	Continuous	✓	Deterministic	×
Xu *et al.* [27]	Demand Response	DDPG	Continuous	✓	Deterministic	×
Duan *et al.* [24]	Operational Control	DQN, DDPG	Discrete & Continuous	✓	Deterministic	×
Ye *et al.* [26]	Electricity Market	DDPG	Continuous	✓	Deterministic	×
Mocanu *et al.* [28]	Energy Management	DQN, DDPG	Discrete & Continuous	✓	Deterministic	×
Our work	Energy Management	PPO	Continuous	✓	Stochastic	✓

Based on the powerful algorithms, DRL has shown advantageous performance in various decision-making tasks. Trained by a novel combination of supervised learning from human expert games, and reinforcement learning from games of self-play, a 99.8 % winning rate was achieved by AlphaGo [17]. Through the deep convolutional neural network and the partial observation, [18] developed a method to learn policies that map raw image observations directly to robot torques. Without human competitive, intelligent support, deep reinforcement learning was performed on advancing automated game testing [20]. Extending the Bayesian personalized ranking with a neural network, the policy detection had been improved against non-stationary agents [19].

B. DRL-BASED POWER SYSTEM MANAGEMENT

In recent years, some pioneering studies have been implemented related to the applications of DRL in the electrical grid. These applications cover some control and optimization problems in different application fields, including operational control, electricity market, demand response, and energy management. To cope with the emergency control in the power system, a novel adaptive strategy and an open-source platform were designed for the grid control [21]. Reference [25] proposed a deep Q-learning based algorithm for local energy trading to optimize the decision-making processes of prosumers. In [22], a novel two-timescale voltage regulation scheme was introduced for smart grids by coupling deep Q-learning with physics-based optimization. The aforementioned works focused on the employment of DQN for grid optimization. However, DQN can only handle discrete and low-dimensional action spaces, which can not meet the requirement of continuous management.

To support continuous control, several researchers have tried to apply the policy gradient method in the regulation domain. In [23], a model-free method based on DRL with continuous action space was introduced to replace the traditional linear controller for load frequency control. Reference [27] applied the deep deterministic policy gradient algorithm to solve the joint bidding and pricing problems of the load-serving entity. Similarly, the autonomous voltage control strategies were proposed by [24] based on DQN and DDPG to support grid operators in making effective control actions. An online optimization was proposed for building energy management systems [28]. Its learning algorithm combined deep Q-learning and deep policy gradients, both of which were extended to perform multiple control actions simultaneously. Ye *et al.* [26] developed a multi-agent deep reinforcement learning approach based on the deep policy gradient to optimize the strategy of multiple self-interested generation companies.

The comparison of related works is summarized in Table 1. The policy type means the mapping from state space to action space, which can be deterministic (outputting action directly) or stochastic (outputting the probability of choosing an action). Our work is different from the above works in several aspects. Firstly, compared with DQN-based methods [21], [22], [25], our work supports continuous action space and can be used to solve a wide range of control problems. Secondly, different from other works, our work realizes a probabilistic policy. Instead of outputting a deterministic action directly, the policy in our work solves approximates the desired probability distribution of the optimal control signals, enabling better flexibility and robustness as increasing uncertainties and randomness in power systems. Thirdly, the training of our approach is based on the proximal policy optimization (PPO), which is proven to have more stable training convergence properties [31] than other DDPG methods [30].

III. ELECTRIC VEHICLE SYSTEM ENERGY FLOW

The rising sales of electric vehicles boost the coupling of vehicles and the electrical grid. To investigate this hybrid system, the EV feature and interaction channel with the power system are reviewed to reveal the energy flow and state dynamics.

A. ELECTRIC VEHICLE DYNAMICS

In parallel with the control center of the power system, distributed EV aggregators regulate the EV fleets based on their battery states. According to [32], an EV aggregator including a large number of EVs can be modeled as the following first-order transfer function:

$$G_{EV(s)} = \frac{K_i^{EV}}{1 + sT_i^{EV}} \tag{1}$$

where $K_i^{EV}, i = 1, 2, \ldots, N$ are the charging and discharging efficiencies of the i^{th} EV aggregator, and T_i^{EV}, $i = 1, 2, \ldots, N$ are the corresponding time constants of the EV

battery system. For the simplification, the charging and discharging efficiencies are set the same coinciding with [32].

Denoting the power command u_i^{EV} from the aggregation center, the energy dynamic of i^{th} EV aggregator can be described with the consideration of SOC variation [33].

$$\begin{bmatrix} \dot{P}_i^{EV} \\ \Delta\dot{SOC}_i \end{bmatrix} = \begin{bmatrix} \frac{-1}{T_i^{EV}} & 0 \\ \frac{-1}{3600} & 0 \end{bmatrix} \begin{bmatrix} P_i^{EV} \\ \Delta SOC_i \end{bmatrix} + \begin{bmatrix} K_i^{EV} \\ 0 \end{bmatrix} u_i^{EV} \tag{2}$$

Since the battery capacity is usually quantified with the unit of Ah, the time constant $\frac{-1}{3600}$ is involved in (2) to unify the time unit. Through the calculation of SOC increment, the aggregator SOC is then obtained adding on the initial SOC [34].

$$SOC_i = \Delta SOC_i + SOC_i^{initial} \tag{3}$$

The SOC of each aggregator varies with diversified traveling schedules, driver behaviors, and battery devices, which form a heterogeneous system with different SOC values. Besides the SOC, the location of EV and the efficiency of the charging station may also be uncertain. Therefore, the generation of power commands u_i^{EV}, $i = 1, 2, \ldots, N$ should comprehensively take all these factors into account.

B. ELECTRIC GRID DYNAMICS

In addition to the EV dynamics, the growing number of charging stations and the increasing speed of charging/discharging power make the mutual energy transmission possible between aggregated EVs and the electrical grid. Since the absorption and release of EV power directly influence the active power, the grid frequency deviation of Δf is then affected [35].

$$\Delta\dot{f} = -\frac{D}{M}\Delta f + \frac{1}{M}\Delta P_g + \frac{1}{M}\sum_{i=1}^{N}\Delta P_i^{EV} - \frac{1}{M}\Delta P_L \tag{4}$$

where D is the grid damping coefficient, M refers to the system inertia. The real-time power imbalance is determined by the power difference among the turbine ΔP_g, the electric vehicle ΔP_i^{EV}, and the fluctuated load consumption ΔP_L.

With the turbine constant T_c and the mechanical strength ΔP_m, the turbine power ΔP_g is generated to compensate the frequency variation [16]

$$\Delta\dot{P}_g = -\frac{1}{T_c}\Delta P_g + \frac{1}{T_c}\Delta P_m \tag{5}$$

Considering the reheat type of stream turbine [16], the relationship between the valve position ΔX_g and ΔP_m is then derived to reflect the reheat dynamic.

$$\Delta\dot{P}_m = \frac{-F_p}{RT_g}\Delta f - \frac{1}{T_r}\Delta P_m + \frac{T_g - F_p T_r}{T_r T_g}\Delta X_g \tag{6}$$

where T_r is the reheat constant, F_p indicates the fraction of total turbine power. The valve position is determined by the droop feature of turbine with R.

$$\Delta\dot{X}_g = -\frac{1}{RT_g}\Delta f - \frac{1}{T_g}\Delta X_g + \frac{\alpha_0}{T_g}u_{grid} \tag{7}$$

It can be seen from (7) that the command u_{grid} from the grid operator and the allocation ratio α_0 determines the regulation value of the governor. As a practical approach, the area control error (ACE) [32] is usually formulated to present the frequency deviation severity.

$$ACE = \beta\Delta f \tag{8}$$

where β is the bias factor for the frequency-response characteristic. Through the linear proportional integral (PI) controller, the u_{grid} in (7) is constructed to guide the governor output.

$$u_{grid} = -K_P ACE - K_I \int ACE \tag{9}$$

Different from the original electrical system, the participation of EV energies and their SOC constraints would directly influence the frequency variation in (4). Moreover, by (5)-(9), the frequency variation couples with the sequential dynamics in the governor and turbine, which affects the energy flow of EV and grid system.

C. ENERGY REINFORCED MANAGEMENT STRUCTURE

Shown from the (9), the electrical grid has already built mature pattern dealing with the frequency fluctuations, but for the EVs, the formulation of aggregation commands u_i^{EV}, $i = 1, 2, \ldots, N$ still face a number of challenges. On account of the time-varying characteristic of the EV system and the feature of the multiple observation inputs, an energy reinforced management structure is proposed in Fig. 1.

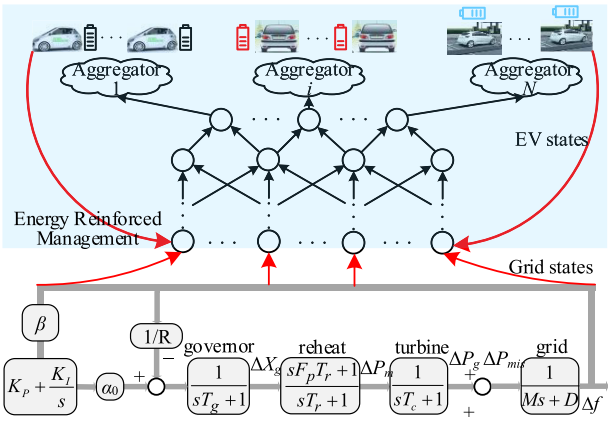

FIGURE 1. Energy reinforced management structure for the EV aggregation.

Considering the multiple state variables Δf, ΔP_g, ΔP_m, ΔX_g, ACE in the electrical grid, as well as the EV dispatch power P_i^{EV}, SOC_i, $i = 1, 2, \ldots, N$, the energy management problem is shown as a multiple-input multiple-output pattern, which is suitable for the deployment of a policy network in Fig. 1. From the grid and EVs, the energy reinforced

management observes the environment variables define as follows.

$$\begin{aligned}\mathbf{S}_{in} &= [\mathbf{S}_{grid}, \mathbf{S}_1^{EV}, \ldots, \mathbf{S}_i^{EV}, \ldots, \mathbf{S}_N^{EV}] \\ &= [\Delta f, \Delta P_g, \Delta P_m, \Delta X_g, ACE, \\ &\quad P_1^{EV}, SOC_1, \ldots, P_N^{EV}, SOC_N] \end{aligned} \tag{10}$$

According to the high-dimensional inputs (10), without disturbing the original automatic generation control ΔX_g, the policy network generates the aggregation commands for EVs with various *SOC* conditions.

$$\mathbf{U}_{EV} = [u_1^{EV}, \ldots, u_i^{EV}, \ldots, u_N^{EV}] \tag{11}$$

By this structure, the dynamics from both electrical grid and EVs can be fully detected by the policy network, and the decision-making advantages of complex network is able to manage the continuous energy variations of EVs. For the electrical grid-electric vehicle system, the main concern is to minimize the frequency deviation defined as:

$$min \, \|\Delta f\| \tag{12}$$

In the meantime, instead of the static commands, the EV SOCs are regulated to make the aggregator with higher SOC provides more power than the aggregator of lower SOC. For instance, the other objective is formulated as:

$$min \sum_{i=1}^{N} \left\| SOC_i - \frac{SOC_1 + \cdots + SOC_N}{N} \right\| \tag{13}$$

Integrating these two objectives (12)-(13), an energy management objective can be generated as an example:

$$min \, \|\Delta f\| + \sum_{i=1}^{N} \left\| SOC_i - \frac{\sum_{i=1}^{N} SOC_i}{N} \right\| \tag{14}$$

As a general structure, it should be mentioned that the inputs and the objective function of the EV energy management can be flexibly changed to meet the different requirements of EV aggregation and grid operation.

IV. ENERGY REINFORCED MANAGEMENT PROCESS

To implement the designed management structure in Section II.C, the establishment of the proposed policy network is further mapped to a Markov decision process to optimize the real-time aggregation commands in (11).

A. MARKOV DECISION PROCESS FOR THE EV MANAGEMENT

In our proposed energy reinforced management strategy, an agent learns how to achieve the control objectives by interacting with the power grid through the control commands. The control problem could be modeled as a Markov decision process (MDP), which provides a common framework for the sequential decision-making process. Normally, the MDP is formulated as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the transition probability, $r(s, a) \in \mathbb{R}$ is the reward function and $\gamma \in [0, 1)$ is the discount factor. In this paper, we consider the frequency control problem of EV charging as a partially observable MDP. According to the energy reinforced management structure in Section II, The components of agents, states, actions, and rewards in the MDP are defined as follows.

Agents. To leverage the advances of artificial intelligence, we model the EV aggregators as intelligent agents who can learn an optimal control policy for the unified management of electric vehicles and the electrical grid. The agent regulates the EV fleets of the system based on the battery states as well as the observable grid metrics.

States. The time-varying environmental states $\mathbf{S}_{in}^{EV}$ defined in (10) are imported to describe the internal agent dynamics. The states of the agents are composed of the multiple state variables $\mathbf{S}_{grid}$ in the electrical grid and the power, energy states $\mathbf{S}_1^{EV}, \ldots, \mathbf{S}_i^{EV}, \ldots, \mathbf{S}_N^{EV}$ of EVs.

Actions. To enhance the robustness of the management strategy, the agents formulate the probability distribution of optimal dispatch action instead of generating deterministic control signals. The policy network outputs the distributions of optimal actions, and actual actions are then sampled from the established distribution, which bids the power commands $u_i^{EV}, i = 1, \ldots, i, \ldots, N$ of different EVs.

Rewards. To participate in the auxiliary frequency service without disturbing the customized energy requirement of EVs, the opposite of energy management objective in (14) is deployed as a rewarding example, which can also adjust according to the real-time scenario.

According to the policy π, the frequency aggregation agent receives the observation s_t and chooses an action $a_t \sim \pi(s_t)$ at each time t, which can be represented as the stochastic policy $\pi(a|s) \sim [0, 1]$ outputting the possibility distribution of actions. The reward function is an incentive measurement to guide the agent to achieve its goal. The immediate reward of a_t at state s_t is noted as $r(s_t, a_t)$ and $R_t = \sum_{i=0\ldots\infty} \gamma^i r_{t+i}$ indicates the cumulative reward.

B. PROBABILISTIC MANAGEMENT FOR ROBUSTNESS ENHANCEMENT

Like most real-world control tasks, EV energy management is also a continuous control problem, which is generally more challenging. Previous work [36] straightforwardly apply Deep Q-Network (DQN) by leveraging discretization over the continuous action space to bypass this challenge. However, the discrete action space cannot provide precise control actions as the control system needs fine-grained discretization. For example, Yang *et al.* [36] configure the EV charging actions as $[-100\%, -50\%, 0\%, 50\%, 100\%]$ with respect to the maximum charging rate, and it is not possible to set the charging rate to 25%. Furthermore, the optimization over the large discrete action space is impractical as the number of actions increases exponentially with the number of control signals. It is impossible to straightforwardly apply

value-based reinforcemelearning (e.g., Deep Q-Network, DQN) in the continuous action space because there is an infinite number of actions to estimate the values, which requires expensive computational effort.

To support continuous control, we introduce a policy network to model and optimize the policy directly. Specifically, the proposed algorithm maintains a parameterized policy function $\pi_\theta(s_t)$, which specifies the control policy. To learn more flexible and intelligent control strategies, the actor function in our work learns the distribution of optimal control instead of outputting the desired control signals directly, which establishes a probabilistic management approach to enhance the performance robustness. Here, we deploy a Gaussian policy as an instance, which is defined as:

$$\pi_\theta(a|s) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{(a-\mu)^2}{2\sigma^2}) \tag{15}$$

where the means and standard deviations are the outputs of $\pi_\theta(s_t)$, which is the policy network in our implementation. The policy network approximates the distribution, and actual control signals are sampled from this distribution.

C. PROXIMAL POLICY OPTIMIZATION FOR THE MANAGEMENT GRADIENT

We follow the policy gradient method to optimize our policy network. The algorithm tries to maximize the expected return $J(\pi_\theta)$ by leveraging the gradient concerning the parameters of its policy π_θ:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{a\sim\pi_\theta}[\nabla_\theta \log(\pi_\theta(a_t|s_t))\hat{A}(s_t, a_t)] \tag{16}$$

where $\hat{A}(s_t, a_t)$ is an estimator of the advantage function at the timestep t. $\hat{A}(s_t, a_t)$ represents the advantage of taking an action a_t at a given state s_t.

$$\hat{A}(s_t, a_t) = R_t - V(s_t) \tag{17}$$

where R_t denotes the return received by the aggregator after running a particular trajectory starting from s_t at time step t, $V(s_t)$ is a value function which is used to estimate the return of starting in s_t and following the policy for subsequent steps.

$$V(s_t) = \mathbb{E}[R_t|\pi_\theta, s_t] \tag{18}$$

The value function $V(s_t)$ defined in (18) is estimated to calculated the advantage function. Following the calculation of the policy gradient defined in Equation (16), the parameter update of π_θ depends on the advantage function defined in Equation (17). The policy gradient method can be interpreted as trying to maximize the likelihood of actions that can result in a higher expected return and to minimize the likelihood of actions that can result in lower expected returns. However, in practice, the estimation of policy gradient suffers from the high variance, which may result in the algorithm instability during training, i.e., the performance may vary dramatically between different iterations. Although the problem caused by the noise gradient can be alleviated by using a large amount of data with each update, the policy gradient algorithm might still be unstable. To address this issue, the proximal policy optimization measures a probability ratio between old and new policies which is defined as follows.

$$w(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} \tag{19}$$

A constraint is then imposed by forcing the ratio to stay within a small interval $[1-\varepsilon, 1+\varepsilon]$ as follows.

$$\min\left(w(\theta)\hat{A}(s,a), \text{clip}(w(\theta), 1-\epsilon, 1+\epsilon)\hat{A}(s,a)\right) \tag{20}$$

where ε is a hyperparameter that controls the clipping degree.

The likelihood radio $w(\theta)$ can be interpreted as a measure of the similarity between the current policy π_θ and the old policy π_{old}. The clip function is used to discourage the policies from deviating too far apart. Therefore, the goal of the algorithm to improve the stability of policy gradient algorithms.

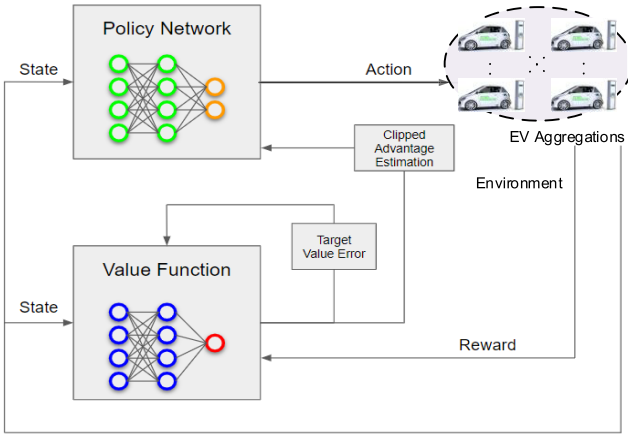

FIGURE 2. Policy network and value function network for the EV aggregation.

The overall learning process is shown in Fig. 2. We adopt two neuron networks for approximating the policy and the value function, respectively. The policy network observes environment states and generates control signals. The value function receives the reward signal and calculates target value error to update its parameters. With the help of the value function, the advantage function can be obtained by following Equation (17). After that, the clipped advantage estimation can be calculated by following Equation (20) to update the parameters of the policy network for each update step.

As shown in Algorithm 1, the goal of our approach is to learn the parameter θ of the optimized policy network π^*. We first initialize the parameters for policy and value functions. At each iteration of training, the agent first collects the required information of a trajectory by running for an episode of N steps. For each step, the agent observes the current state s_t and calculates the output of the policy network. In our case, the output of the network is the mean and variance of the action distribution defined in (15). The control signals are sampled from this distribution. The historical information of one trajectory is stored in the

Algorithm 1 Energy Reinforcement Management of EV Aggregation for the Frequency Regulation

Input: the episode number L;
the maximum step number N of each episode;
Output: optimal control policy π_θ^*
Initialize the policy network: $\theta \leftarrow$ random weights;
Initialize the value function: $\psi \leftarrow$ random weights;
Set up the hyperparametes for network training;
Initialize the trajectory storage $\mathcal{H}$;
for *episode* $= 1, 2, \ldots, L$ **do**
  Initialize state s_0 for the EV and grid system;
  for $t = 1, 2, \ldots, N$ **do**
    Calculate the output of the actor: $(\mu, \sigma) = \pi_\theta(s_t)$;
    Sample control signal $a_t \sim \mathcal{N}(\mu, \sigma)$;
    Perform the control signal a_t and receive reward r_t, new state s_{t+1} and finish signal *done*
    Store transition (s_t, a_t, r_t, s_{t+1}) in $\mathcal{H}$;
    if *done* **then**
      break;
  end for
  $\theta_{old} \leftarrow \theta$;
  for *each update step* **do**
    Sample minibatch of D samples $\{(s_i, a_i, r_i, s_i')\}$ from $\mathcal{H}$;
    // Update the value function
    for (s_i, a_i, r_i, s_i') **do**
      $y_i \leftarrow$ compute target values;
    end for
    $\psi \leftarrow \psi + \alpha_v(\frac{1}{n}\sum_i \nabla_\psi V_\psi(s_i)(y_i - V(s_i)))$
    // Update the actor function
    for (s_i, a_i, r_i, s_i') **do**
      $\hat{A}_i \leftarrow$ compute advantage using V_ψ;
      $w_i(\theta) \leftarrow \frac{\pi_\theta(a_i|s_i)}{\pi_{\theta_{old}}(a_i|s_i)}$;
    end for
    $\theta \leftarrow \theta + \alpha_\pi \frac{1}{n}\sum_i \nabla_\theta min(w_i(\theta)\hat{A}_i, clip(w_i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i)$
  end for
end for

trajectory storage $\mathcal{H}$. The training starts at the end of each episode. For each update step of training, the algorithm first samples a batch of D samples from $\mathcal{H}$. The target values y_i are used to update ψ with a stepsize of α_v. For each sample in $\{(s_i, a_i, r_i, s_i')\}$, the advantage $\hat{A}_i$ and the probability radio w defined in (19) is calculated. The parameter θ of policy network is updated following (20) with a stepsize of α_π. Finally, as the training goes on, the probabilistic dispatch center π_θ gradually converges to a robust control policy. The complete implementation process for the energy reinforcement learning of EV aggregation is displayed in Algorithm 1.

V. MANAGEMENT CASE DEMONSTRATION

To demonstrate the performance of the proposed method, a 5GWh electrical grid with two EV aggregators is employed here as an example. The same structure of test system is adopted in [16], [32], [35].

A. CONFIGURATION FOR THE SYSTEM TEST

1) PARAMETERS OF ELECTRICAL GRID AND EVs

The grid inertia M and damping D are set as 8.8 and 1, respectively. With the time constants of governor T_g=0.2, reheat T_r=12, turbine T_c=0.3, and turbine fraction Fp=1/6, the governor power is transferred into the grid to maintain the system frequency. According to the Tesla V3 supercharging, the power range of each EV is assumed to vary between [−250kW, +250kW]. The parameters in Table 2 for the EV aggregators are decided according to [32]. The charging/discharging coefficients K_1^{EV}, K_2^{EV} present the battery efficiency, while the time constants T_1^{EV}, T_2^{EV} are the indicators of battery response speed. The participation rates α_1, α_2 of aggregators are determined based on the energy contribution ratio between the electrical grid and the EV aggregator. In this study, the energy capacity of the electrical grid is 5GWh, while the EV Aggregator 1 contributes 0.125 GWh in the frequency regulation service and the contribution from EV Aggregator 2 is 1.125 GWh. Hence, the participation ratio α_1 = 0.125GWh/(5GWh+0.125GWh+1.125GWh)=0.02, α_2 = 1.125GWh/(5GWh+0.125GWh+1.125GWh)=0.18. The initial SOC of Aggregator 1 is 48%, while Aggregator 2 starts with SOC_2=52%. If 10000 EVs charge simultaneously, the peak power is able to reach 10000*250kW=2.5GW, which forms a relatively large energy integrated system.

TABLE 2. Electric vehicle aggregator parameters.

Parameters	Symbol	Values
Charging/discharging coefficient	K_1^{EV}, K_2^{EV}	1
Battery time constant	T_1^{EV}, T_2^{EV}	0.1
EV participation rate	α_0	0.2
Aggregator 1 participation rate	α_1	0.02
Aggregator 1 EV initial SOC	SOC_1	0.48
Aggregator 1 EV number	N_1	1000
Aggregator 2 participation rate	α_2	0.18
Aggregator 2 EV initial SOC	SOC_2	0.52
Aggregator 2 EV number	N_2	9000

According to [16], the initial SOCs are randomly set to verify the management performance of the proposed strategy. The SOC changes of EVs are determined by the power variation of EV aggregators. The changes of EV SoCs are closely related with the power demand, systems disturbances, initial SOC, configuration, renewable energy, and charging/discharging time of EVs. All these uncertain parameters can be classified into the external and internal variations demonstrated in Section VC and Section VD of the manuscript. In terms of the external disturbance, the PV measurement data [37], system disturbance of power increase are investigated. For the internal disturbance, the initial SOC and system configuration of the participation ratio is taken

into account. Through the lumped parameters T_1^{EV}, T_2^{EV}, the charging/discharging time of EV aggregators is also included.

In the real-time application, the dynamics of power systems and EVs are complex and nonlinear. However, the electrical grid and electric vehicles are integrated through the secondary frequency service, which operates in the time-scale of 1-10 minutes in this study [38]. During this time scale, the frequency regulation of the power system can be approximately treated as a linear model [16]. Meanwhile, instead of an individual EV, the EV aggregator with many EVs is considered to participate in the auxiliary service as a whole, which can also be simplified as a linear model [32], [35]. Moreover, even for each battery, the power output in the short time of 1-10 min is also the same as the power commend. Therefore, through the power variation in the electrical grid-electric vehicle system [16], [32], [35], the EV SOC is analysed in this study. On account of the renewable energy, the PV measurement data on June 17, 2012 from Electric Power Research Institute [38] is adopted as the external power disturbance to verify the management effectiveness.

2) PARAMETERS OF DESIGNED MANAGEMENT SYSTEM

Following paper [31], the discount factor γ for our training algorithm is set to 0.99, and the clip parameter ϵ of PPO is 0.2. There are two dense layers for both policy network and value network, and each layer with 64 neurons. The learning rate of policy and value network for PPO is set as 0.0003 and is decayed linearly with the learning episode. The value network outputs expected accumulated rewards of the current input state. The output layer of the policy network is the Gaussian distribution (e.g., the outputs are mean μ and variance σ) for continuous control. The output actions are stochastic and sampled from these distributions.

The proposed reinforcement learning algorithms and simulation environment are developed by Python. These programs run on a server with 12-core Intel(R) Xeon(R) CPU E5-1650 v3 @ 3.50GHz processors. The deep neuron network training of deep reinforcement learning is accelerated on 1 NVIDIA GTX 1080Ti GPU.

B. MANAGEMENT EVOLUTION

For the EV management system, the discount factor γ of deep reinforcement learning is 0.99. The learning stepsizes α_π and α_v for the policy network and value function network are 0.0002 and 0.0005, respectively. The episode number is set as 2000 with the maximum step number of 20000. To guarantee the algorithm stability, the ratio between old and new policies is clipped via $\epsilon = 0.2$. Injecting a small load disturbance ΔP=0.1p.u, the training process is shown in Fig. 3, which demonstrates three progressive stages: adaption, exploration, and stabilization. The red and blue curves in Fig. 3 compare the rewards of the proposed energy reinforced management and the conventional linear dispatch without energy reinforced management. Between 0-382 episodes, i.e., the

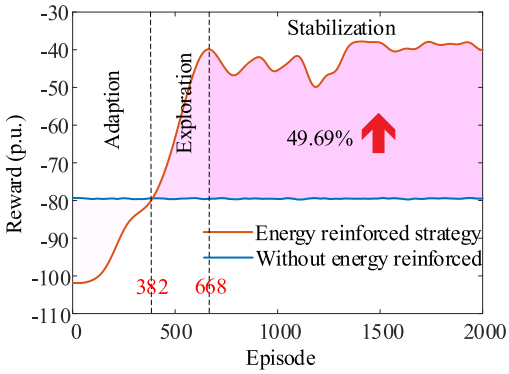

FIGURE 3. Episode performance improvement during the training procedure.

adaption stage, the reinforcement strategy gradually upgrades its performance from -101.86 to -80.06, which reaches the same reward of traditional linear strategy. However, because of its evolutionary mechanism, the developed strategy is able to continue the reward progress until 668^{th}, which empowers the system with 49.69% enhancement in the exploration stage. After that, the whole management system steps into the stabilization stage, maintaining the reward around -40 p.u.. Despite the weak regulation performance at the beginning, the management network can effectively find the optimal dispatch outputs for the multiple EV aggregators within only 908 episodes during the adversary against the value function network.

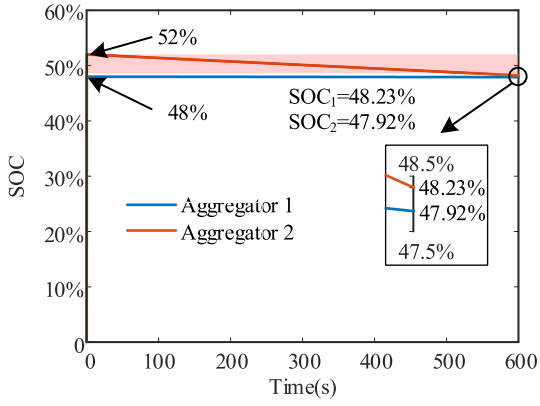

FIGURE 4. SOC management with the developed energy reinforced approach.

Fig. 4 shows that the SOCs of Aggregator 1 and 2 almost converge within 600s. During this period, the SOC of aggregator one declines from 48% to 47.92%, while the Aggregator 2 SOC rapidly drops from 52% to 48.23%. As indicated by Fig. 5, the frequency deviation is maintained in the arrange of [-0.013Hz, 0.013Hz], which secures the normal frequency state of the electrical grid. From the power outputs of EVs and the turbine in Fig. 6, it is seen that facing the external disturbance and the internal energy imbalance, the whole system adapts in the 50s. The stable power from EV

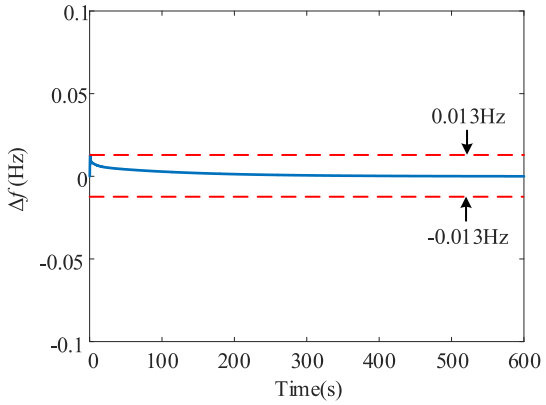

FIGURE 5. Grid frequency facing the load disturbance.

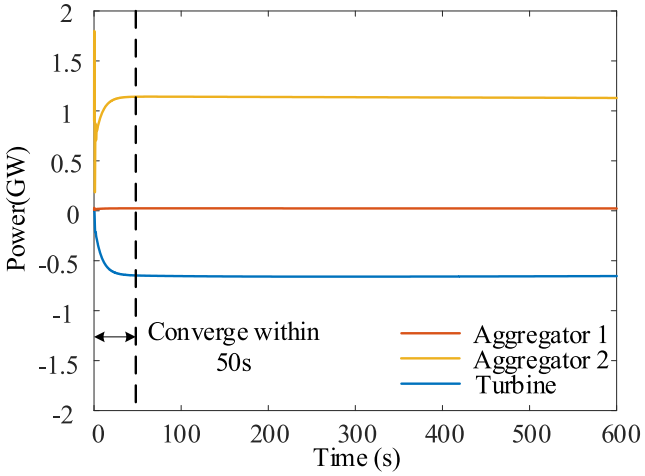

FIGURE 6. Power outputs under the reinforced management.

Aggregator 1 is 0.024GW, while the Aggregator 2 generates 1.13GW to compensate for the external fluctuation and decrease their SOC difference. In the 50s, the gradual optimization process is also displayed in Fig. 6. In the beginning, the power from Aggregator 2 quickly rises to 1.80GW. The occurrence of this quick rise is caused by the management objective and mechanism of reinforcement learning. As the management target is set to compress the frequency fluctuation and SOC imbalance simultaneously, the energy management system aims to explore the optimal operation point facing the internal and external disturbances. Through the rapid exploration in the range of [0, 1.80GW], it helps the proposed policy find out the final optimal point of 1.13GW. After that, aggregated EVs slowly decrease the power outputs converging to their optimal operating states. The EV aggregator with a higher SOC level participates more than the aggregator with lower SOC value.

C. REINFORCED MANAGEMENT AGAINST EXTERNAL VARIATIONS

Considering the external variations, the power disturbance injecting to the system increases from 0.1 p.u. to 0.2 p.u.. Facing this change, the SOC changes of the two EV aggregators are listed in Table 3. Due to the larger disturbance, the reinforcement-based management dispatches more power from the Aggregator 2. Within 500s, the SOC difference between the two EV groups is almost eliminated by the proposed strategy shown in Table 3. Compared with the SOC regulation speed in Fig. 4, the SOCs of two aggregators coincide within 500s, which reflects the adaption of the proposed strategy against external variations. The corresponding power curves from the two aggregators are depicted in Fig. 7. It can be seen that the power output from the Aggregator 2 has grown to 1.5GW to reduce the external variations, which is 1.27 times larger than that in Fig. 6. Without knowing the system model and disturbance, the developed strategy can automatically change, guaranteeing the system flexibility.

TABLE 3. Rapid energy balance among different aggregation groups.

Time	Aggregator 1 SOC	Aggregator 2 SOC
0-500s	$48.00\% - 47.93\%$	$52.00\% - 47.98\%$

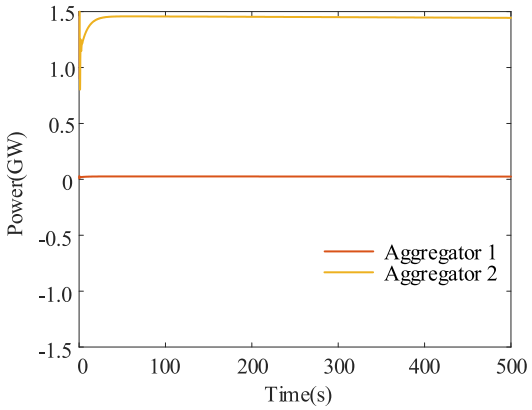

FIGURE 7. Power outputs with the increased disturbance of 0.2 p.u.

On account of the real-time variation, the variation data based on the PV measurement [37] is deployed to vary the electrical grid. In 600s of Fig. 8, the power disturbance fluctuates between 0.342GW and 0.805GW, which reaches 16% of the power rating. With the help of deep reinforcement learning, the frequency deviation is displayed in Fig. 9. Facing the high-penetration of renewable energy, the frequency of the whole system only varies between [−0.015Hz, +0.015Hz], which demonstrates the feasibility of the proposed management. From Fig. 9, it can also be seen that the frequency deviation curve is much smoother after 300s showing the continuous learning capability of the proposed management strategy.

D. REINFORCED MANAGEMENT AGAINST EV INTERNAL VARIATIONS

Besides the external disturbance, the states of EV aggregators may also suffer from internal dynamics and uncertainties.

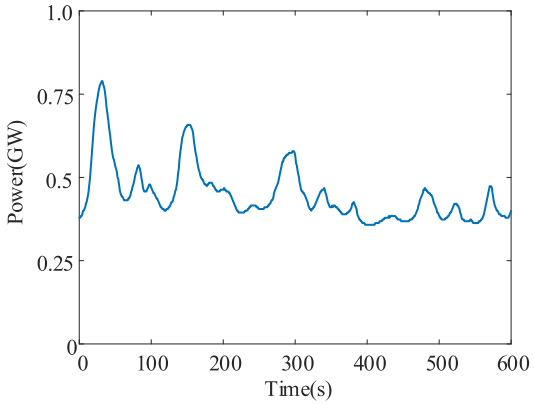

FIGURE 8. The PV fluctuation in 10 minutes.

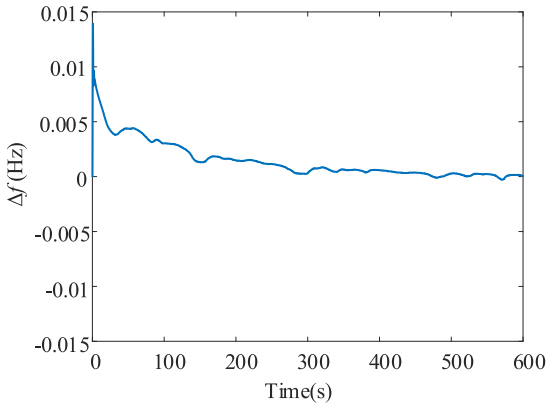

FIGURE 9. Grid frequency deviation facing the PV fluctuation.

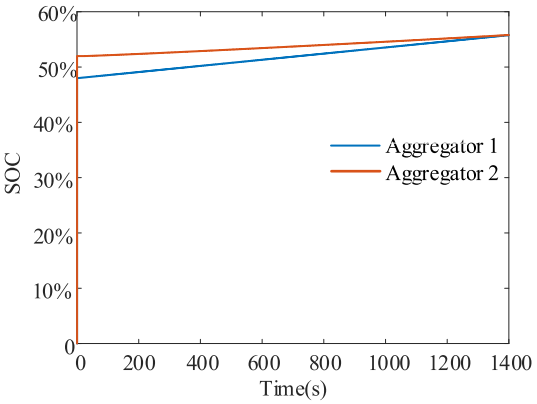

FIGURE 10. SOC adjustment with internal aggregation uncertainties.

To verify the advantage of developed management strategy, the participation rates of two EV aggregators are switched from 1:9 to 1:1, which means these two aggregators have an identical contribution to the frequency regulation. The SOC adjustment process is shown in Fig. 10. With higher power, the EV Aggregator 1 increases its SOC from 48% to 55.77 %, while the Aggregator 2 slowly grows to 55.80 %, which keeps the dynamic balance via the model-free dispatch. To investigate the dispatch process of

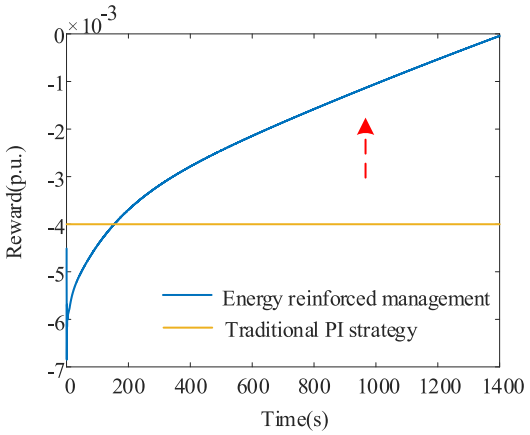

FIGURE 11. The reward comparison with traditional PI AGC signals.

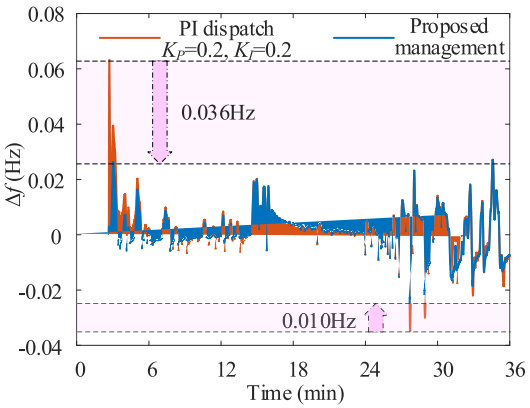

FIGURE 12. Performance consistency against internal and external dynamics.

the deep reinforced management, the reward of each step is also displayed in Fig. 11. Although there is a small adaption stage in the period of [0s, 155s], the objective is gradually optimized to 0 by the developed reinforced management. However, the traditional automatic generation control signal from the proportional-integral (PI) controller is unable to adjust the SOC dynamically (K_P=0.2, K_I=0.2), which influences the aggregation performance of the EV system.

To further detect the management performance facing various SOC scenarios, the SOC values of two EV aggregators are set the same with 52%. Meanwhile, the timescale of external PV fluctuation is extended to 36 minutes for the verification of long-term performance consistency. From the long-term simulation in Fig. 11, the peak value of the frequency deviation is constrained by 0.036Hz, while the minimum frequency dip is reduced by 0.01Hz. Defining the comparison index J as the deviation reduction per hour, its expression is as below

$$J = \sum_{t}^{t+\Delta T} \|\Delta f\| / \Delta T \tag{21}$$

According to (21), the accumulated frequency deviation of PI dispatch is 185.38Hz/hour, while the proposed management is 163.82Hz/hour. In front of this random dynamic, the improvement can still reach 11.63%. As time goes by, the merit of an established management strategy will be more notable, which makes the intelligent EV management possible.

VI. CONCLUSION

A probabilistic energy reinforced management structure is developed in this work to cope with the modeling and aggregation difficulties of the electric vehicles (EVs). Parallel with the traditional automatic generation control, the policy network in the proposed management strategy observes the operating dynamics of the electrical grid and the electric vehicles simultaneously to fully understand and utilize the energy differences. According to the instantaneous and the accumulated rewards from the Markov decision process, the established policy network continuously generates and adjusts its dispatch commands for various EV aggregators. To enhance the algorithm consistency, the policy network gradient is further optimized by regulating the evolutionary ratio between the old and new policies. By the typical electrical grid-electric system, the proposed approach demonstrates significant merits in terms of SOC dispatch, frequency regulation, and the evolutionary properties, which establishes a uniform energy management framework for the aggregation of EVs in the grid frequency regulation. The future work is to take the aggregation process into consideration. In this study, the EV aggregation process is simplified based on references, but the practical aggregation dynamic is complex and requires the investigation of diversified customer behaviors. As the proposed management strategy does not rely on the system model and includes probability learning, it could be further designed and extended to the hybrid energy-behavior management for effective EV-grid integration.

REFERENCES

[1] The Guardian. *Tesla Delivers First China-Made Cars From 5BN Shanghai Factory*. Accessed: 2019. [Online]. Available: https://www.theguardian.com/technology/2019/dec/30/tesla-delivers-first-china-made-cars-from-shanghai-factory

[2] Forbes. *How is Musk Going to Pay For Tesla's Chinese Gigafactory*. Accessed: 2018. [Online]. Available: https://www.forbes.com/sites/alanohnsman/2018/07/10/how-is-musk-going-to-pay-for-teslas-chinese-gigafactory/#29ba49505739

[3] CNN Business. *Tesla Starts Building Its Huge Shanghai Factory to Make Cars for China*. Accessed: 2019. [Online]. Available: https://edition.cnn.com/2019/01/07/business/elon-musk-tesla-gigafactory/index.html

[4] NIO News. *Believe in Better Nio Day 2019 Held in Shenzhen*. Accessed: 2019. [Online]. Available: https://www.nio.com/news/believe-better-nio-day-2019-held-shenzhen

[5] Bloomberg Businessweek. *The World's Biggest Electric Vehicle Company Looks Nothing Like Tesla*. Accessed: 2019. [Online]. Available: https://www.bloomberg.com/news/features/2019-04-16/the-world-s-biggest-electric-vehicle-company-looks-nothing-like-tesla

[6] Bloomberg Businessweek. *The World'S Biggest Electric Vehicle Company Looks Nothing Like Tesla*. Accessed: 2019. [Online]. Available: https://www.bloomberg.com/news/features/2019-04-16/the-world-s-biggest-electric-vehicle-company-looks-nothing-like-tesla

[7] Tesla. *Introducing V3 Supercharging*. Accessed: 2019. [Online]. Available: https://www.tesla.com/blog/introducing-v3-supercharging

[8] NIO. *Nio Supercharging Pile*. Accessed: 2019. [Online]. Available: https://app.nio.com/content/503398?content_type=article&&content_id=503398&_viewBeginTop=true&load_js_bridge=true&show_navigator=false

[9] BYD. *Byd Charging Services*. Accessed: 2019. [Online]. Available: https://sg.byd.com/page-services-change-services

[10] J. de Hoog, T. Alpcan, M. Brazil, D. A. Thomas, and I. Mareels, "A market mechanism for electric vehicle charging under network constraints," *IEEE Trans. Smart Grid*, vol. 7, no. 2, pp. 827–836, Mar. 2016.

[11] E. L. Karfopoulos, K. A. Panourgias, and N. D. Hatziargyriou, "Distributed coordination of electric vehicles providing V2G regulation services," *IEEE Trans. Power Syst.*, vol. 31, no. 4, pp. 2834–2846, Jul. 2016.

[12] H. Yang, S. Zhang, J. Qiu, D. Qiu, M. Lai, and Z. Dong, "CVaR-constrained optimal bidding of electric vehicle aggregators in day-ahead and real-time markets," *IEEE Trans. Ind. Informat.*, vol. 13, no. 5, pp. 2555–2565, Oct. 2017.

[13] X. Chen and K.-C. Leung, "Non-cooperative and cooperative optimization of scheduling with Vehicle-to-Grid regulation services," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 114–130, Jan. 2020.

[14] T. N. Pham, S. Nahavandi, L. V. Hien, H. Trinh, and K. P. Wong, "Static output feedback frequency stabilization of time-delay power systems with coordinated electric vehicles state of charge control," *IEEE Trans. Power Syst.*, vol. 32, no. 5, pp. 3862–3874, Sep. 2017.

[15] S. Debbarma and A. Dutta, "Utilizing electric vehicles for LFC in restructured power systems using fractional order controller," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 2554–2564, Nov. 2017.

[16] S. Falahati, S. A. Taher, and M. Shahidehpour, "Grid secondary frequency control by optimized fuzzy control of electric vehicles," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 5613–5621, Nov. 2018.

[17] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, p. 484, 2016.

[18] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1334–1373, 2016.

[19] Y. Zheng, Z. Meng, J. Hao, Z. Zhang, T. Yang, and C. Fan, "A deep Bayesian policy reuse approach against non-stationary agents," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 954–964.

[20] Y. Zheng, C. Fan, X. Xie, T. Su, L. Ma, J. Hao, Z. Meng, Y. Liu, R. Shen, and Y. Chen, "Wuji: Automatic online combat game testing using evolutionary deep reinforcement learning," in *Proc. 34th IEEE/ACM Int. Conf. Automated Softw. Eng. (ASE)*, Nov. 2019, pp. 772–784.

[21] Q. Huang, R. Huang, W. Hao, J. Tan, R. Fan, and Z. Huang, "Adaptive power system emergency control using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1171–1182, Mar. 2020.

[22] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-timescale voltage control in distribution grids using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2313–2323, May 2020.

[23] Z. Yan and Y. Xu, "Data-driven load frequency control for stochastic power systems: A deep reinforcement learning method with continuous action search," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1653–1656, Mar. 2019.

[24] J. Duan, D. Shi, R. Diao, H. Li, Z. Wang, B. Zhang, D. Bian, and Z. Yi, "Deep-reinforcement-learning-based autonomous voltage control for power grid operations," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 814–817, Jan. 2020.

[25] T. Chen and W. Su, "Local energy trading behavior modeling with deep reinforcement learning," *IEEE Access*, vol. 6, pp. 62806–62814, 2018.

[26] Y. Ye, D. Qiu, J. Li, and G. Strbac, "Multi-period and multi-spatial equilibrium analysis in imperfect electricity markets: A novel multi-agent deep reinforcement learning approach," *IEEE Access*, vol. 7, pp. 130515–130529, 2019.

[27] H. Xu, H. Sun, D. Nikovski, S. Kitamura, K. Mori, and H. Hashimoto, "Deep reinforcement learning for joint bidding and pricing of load serving entity," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6366–6375, Nov. 2019.
[28] E. Mocanu, D. C. Mocanu, P. H. Nguyen, A. Liotta, M. E. Webber, M. Gibescu, and J. G. Slootweg, "On-line building energy optimization using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3698–3708, Jul. 2019.
[29] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.
[30] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*. [Online]. Available: http://arxiv.org/abs/1509.02971
[31] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*. [Online]. Available: http://arxiv.org/abs/1707.06347
[32] K. S. Ko and D. K. Sung, "The effect of EV aggregators with time-varying delays on the stability of a load frequency control system," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 669–680, Jan. 2018.
[33] Y. Wang, Y. Xu, Y. Tang, K. Liao, M. H. Syed, E. Guillo-Sansano, and G. M. Burt, "Aggregated energy storage for power system frequency control: A finite-time consensus approach," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3675–3686, Jul. 2019.
[34] P. Shen, M. Ouyang, L. Lu, J. Li, and X. Feng, "The co-estimation of state of charge, state of health, and state of function for lithium-ion batteries in electric vehicles," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 92–103, Jan. 2018.
[35] H. Jia, X. Li, Y. Mu, C. Xu, Y. Jiang, X. Yu, J. Wu, and C. Dong, "Coordinated control for EV aggregators and power plants in frequency regulation considering time-varying delays," *Appl. Energy*, vol. 210, pp. 1363–1376, Jan. 2018.
[36] Y. Yang, J. Hao, Y. Zheng, and C. Yu, "Large-scale home energy management using entropy-based collective multiagent deep reinforcement learning framework," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 630–636.
[37] Electric Power Research Institute. *Distributed PV Monitoring and Feeder Analysis*. Accessed: 2019. [Online]. Available: https://dpv.epri.com/#targetText=EPRI
[38] PJM. *Primary Frequency Response Stakeholder Education*. Accessed: 2017. [Online]. Available: https://www.pjm.com/~/media/committees-groups/task-forces/pfrstf/20170901/20170901-primary-frequency-response-education-part-1-of-2.ashx

CHAOYU DONG received the B.Sc. degree in electrical engineering from Tianjin University, Tianjin, China, in 2013, where he is currently pursuing the Ph.D. degree in electrical engineering.

He is also with the Energy Research Institute, Nanyang Technological University. His research interests include power electronics, power system, and energy-information stability analysis and management.

JIANWEN SUN received the B.E. and Ph.D. degrees from Beihang University, Beijing, China, in 2011 and 2018, respectively.

He is currently a Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University. His research interests include deep reinforcement learning, multiagent systems, intelligent smart grid control algorithms, power electronics, fault diagnosis, machine learning, data mining, and deep learning.

FENG WU (Member, IEEE) was born in Hubei, China, in 1990. He received the B.E. and Ph.D. degrees from the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2013 and 2017, respectively.

From 2017 to 2019, he was a full time Senior Engineer at Huawei Technologies Company, Ltd., Shenzhen, China, working on machine learning and artificial intelligence algorithms, such as objection detection and image classification. Since June 2019, he has been a Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University. His research interests include power electronics, fault diagnosis, machine learning, data mining, and deep learning.

HONGJIE JIA (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Tianjin University, China, in 2001.

He became an Associate Professor with Tianjin University, in 2002, where he was promoted as a Professor, in 2006. His research interests include power reliability assessment, stability analysis and control, distribution network planning and automation, and smart grids.