An Improved Two-Stage Deep Reinforcement Learning Approach for Regulation Service Disaggregation in a Virtual Power Plant

Zhongkai Yi, *Member, IEEE*, Yinliang Xu, *Senior Member, IEEE*, Xue Wang,
Wei Gu, *Senior Member, IEEE*, Hongbin Sun, *Fellow, IEEE*, Qiuwei Wu, *Senior Member, IEEE*,
and Chenyu Wu, *Member, IEEE*

***Abstract*—Managing numerous distributed energy resources (DERs) within the virtual power plant (VPP) is challenging due to inaccurate parameters and unknown dynamic characteristics. To address these obstacles, a two-stage deep reinforcement learning approach is proposed for the VPP to provide frequency regulation services and issue the disaggregation commands to DER aggregators in real-time operation. In the offline-stage, an offline simulator is formulated to learn the dynamic characteristics of DER aggregators, through which the soft actor-critic (SAC) algorithm is employed to train the control policy. In the online-stage, the trained control policy is updated continuously in the practical environment, which can ameliorate the performance of the start-up process with prior knowledge. Moreover, a novel sharpness-aware minimization based soft actor-critic (SAM-SAC) algorithm is proposed to improve the robustness and adaptability of the deep reinforcement learning approach. Simulation results illustrate that the proposed approach enables the VPP to manage the DER aggregators to track the regulation requests more accurately and economically than the state-of-the-art methods.**

***Index Terms*—Virtual power plant, regulation service, disaggregation, deep reinforcement learning, sharpness aware minimization.**

Nomenclature

B, b	Mileage cost coefficient of aggregator and DER.
C, c	Incremental operational cost coefficients of aggregator and DER.
E, e	Actual residual energy of aggregator and DER.
$\tilde{E}$, $\tilde{e}$	Intraday residual energy dispatch plans of the aggregator and DER.
G	Gain coefficient.
H	Inertia coefficient.
P, p	Actual power output of aggregator and DER.
$\tilde{P}$, $\tilde{p}$	Intraday active power dispatch plans of aggregator and DER.
ΔP, Δp	Incremental power output of aggregator and DER.
T	Time delay coefficient.
Δt	Control time interval.
U, u	Reference power output to aggregator and DER.
ΔU, Δu	Regulation command to aggregator and DER.
ξ	Energy dissipation rate.
η^{in}, η^{out}	Charging and discharging coefficients.

I. Introduction

THE GROWING number of distributed energy resources (DERs) possess great potential benefits to provide extra flexibility for the power systems [1]. However, limited by the minimum admittance capacity (typically from 1 to 10 MW) issued by the power system operator, individual small capacity DERs cannot directly participate in the superior power system operation [2], [3]. In light of this, the virtual power plant (VPP) emerges as a new platform to realize the coordination between the power system operator and the individual DER devices [4].

VPP has been extensively studied in the power system operation in recent literature, such as unit commitment [5], economic dispatch [6], energy arbitrage [7], etc. It has been demonstrated that the VPP technology can promote the efficient management of various DERs and exploit the complementarity of heterogeneous devices [5], [6]. Most of the existing researches focus on the VPP operations at the time-scale of one hour or fifteen minutes interval [5]–[7]. However, many VPPs contain DERs with rapid regulation capability, which have great potentials to

Manuscript received October 7, 2021; revised January 25, 2022 and March 16, 2022; accepted March 19, 2022. Date of publication March 29, 2022; date of current version June 21, 2022. This work was supported in part by the Shenzhen Science and Technology Program under Grant JCYJ20210324130811031; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515012450; and in part by the Tsinghua Shenzhen International Graduate School Interdisciplinary Research and Innovation Fund under Grant JC2021004. Paper no. TSG-01626-2021. *(Corresponding author: Yinliang Xu.)*

Zhongkai Yi is with the Tsinghua–Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China, and also with the Alibaba DAMO Academy, Alibaba Group, Beijing 100020, China (e-mail: yzk_article@163.com).

Yinliang Xu and Qiuwei Wu are with the Tsinghua–Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China (e-mail: xu.yinliang@sz.tsinghua.edu.cn).

Xue Wang is with the Alibaba DAMO Academy, Alibaba Group, Beijing 100020, China.

Wei Gu and Chenyu Wu are with the School of Electrical Engineering, Southeast University, Nanjing 210096, China.

Hongbin Sun is with the State Key Laboratory of Power Systems, Department of Electrical Engineering, Tsinghua University, Beijing 100084, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSG.2022.3162828.

Digital Object Identifier 10.1109/TSG.2022.3162828

provide extra frequency regulation support for the power systems [8].

Traditional frequency regulation service relies on the primary frequency response and automatic generation control provided by large-scale synchronous generators [9]. The regulation service offering problem for the VPP is much more complicated because it is difficult to model the dynamic characteristic and operational pattern of DER devices. In [10], the primal-dual-gradient method is adopted to solve the real-time regulation problem of the VPP, through which the aggregate DERs are operated like the conventional generators to support automatic generation control or regulation services. The feedback control schemes considering the state-bin transition model of the aggregate DERs are proposed in [11] and [12], which can alleviate the system frequency deviation and energy imbalance.

To coordinate with the prevailing frequency regulation service requirement, it is critical to develop an efficient disaggregation approach for the VPP to track the instantaneous regulation request accurately and economically. Nevertheless, the following challenges with existing literature still need to be further addressed. *First*, the frequency regulation requests issued by the system operator change continuously and frequently (typically every few seconds) [2], [3]. To realize fast disaggregation of the reference regulation request, it is necessary to compute the VPP disaggregation plans in a timely manner. *Second*, the dynamic characteristic of the DERs should be considered in the VPP disaggregation process for accurate and fast response. Most of the existing VPP regulation methods are implemented in a deterministic environment with accurate model parameters [10], [11], [12]. However, the dynamic parameters of the DERs are difficult to calibrate due to the diversified characteristics of different DERs. *Last*, since the regulation command is superposed based on the active power trajectory, the impacts of the regulation service on the intraday dispatch plans should also be considered [13].

Deep reinforcement learning (DRL) is a branch of machine learning methods that focuses on how to make sequential decisions in uncertain environments efficiently. Therefore, the above challenges motivate the utilization of the DRL approach to address the regulation service disaggregation problem in the VPP. Furthermore, the DRL approaches learn the control policy according to the data collected from practical environments or high-fidelity simulators, and make timely decisions without relying on an accurate environment model. DRL has been extensively employed in various power system applications, i.e., frequency regulation [14], market arbitrage [15], energy management [16] and trading [17]. However, there is almost no work studying the VPP disaggregation problem using the DRL approach, and this article could fill the gap in the existing literature. In addition, Liu and Wu [18] propose an appealing two-stage DRL approach, where the offline agent is trained to accumulate knowledge first, and then the online agent improves system safety and training efficiency based on the prior knowledge. However, to address noise in the environment and inevitable discrepancies between the offline and online environments, the DRL method with high adaptability and robustness deserves further investigation.

To address the abovementioned issues, the major contributions and novelties of this article are summarized as follows.

i). This study *firstly advances* the state-of-the-art soft actor-critic algorithm [19], [20] by using the sharpness aware minimization method [21], [22]. Compared with the traditional soft actor-critic, the proposed algorithm can provide higher robustness to the reward noise and is more adaptive to the changes of the environmental parameters.

ii). To the authors' best knowledge, this study for *the first time* addresses the regulation service disaggregation problem in a VPP by using the DRL approach. The proposed approach can efficiently decompose the aggregate regulation request to DER aggregators in the inaccurate environment model.

iii). An offline simulator considering DERs dynamic characteristics is formulated to approximate the practical environment and accumulate experiences for the DRL agent. After that, the trained control policy is updated continuously in the practical environment. The proposed two-stage implementation framework can enhance the tracking accuracy and reduce the expensive operational cost of the *online start-up process* compared with single-stage approaches [14]–[17].

II. Preliminaries

Prerequisite knowledge, including the regulation service disaggregation problem and the mathematical model of DERs, are introduced in this section.

A. Description of Regulation Service Disaggregation Problem

The real-time frequency regulation service disaggregation in a VPP involves three categories of participants: system operator, VPP agent, and controllable units. The system operator refers to the frequency regulation service market operator. VPP agent manages various types of controllable units, including DER aggregators, conventional generators, and renewable energy generations. The relationship and input/output information among the market operator, VPP agent, and controllable units are shown in Fig. 1.

During the real-time operation, the regulation request issued by the system operator changes every four seconds within the VPP regulation capacity bidding range. Therefore, the objective of the VPP is to track the regulation request accurately and economically. The VPP agent needs to efficiently decompose the reference regulation request and issue the regulation commands to all controllable units. Furthermore, the regulation commands issued by the VPP should be corrected dynamically according to the feedback of devices' state sampling.

The relationship between this study and the day-ahead/intraday economic dispatch problem is explained as follows. The real-time regulation services disaggregation problem (with several seconds control interval) is essentially the supplement and expansion of the intraday economic dispatch problem (with one hour or fifteen minutes control interval).

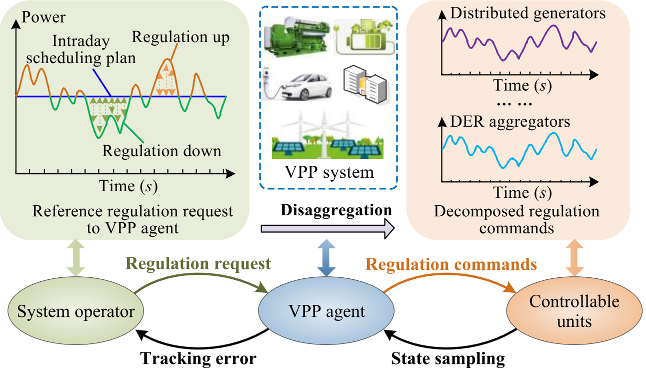


Fig. 1. Relationship among system operator, VPP, and DER aggregators.

During the day-ahead or intraday economic dispatch process, the VPP operator should reserve sufficient frequency regulation capacity and issue the regulation capacity reservation plans for all the controllable units, as in our previous research [23]; and during the real-time operation, the VPP receives the regulation command from the ancillary services market and decomposes the reference regulation command to the controllable units, which is investigated in this study. Thus, the regulation commands to the controllable units are adjusted based on the intraday economic dispatch plan.

B. DER Operation Model

1) Dynamic Characteristic: Considering the communication delay [24] and dynamic inertia [25], the dynamic characteristic of a DER can be generally modelled as follows:

$$\dot{p}_i(\tau) = -\frac{1}{H_i}p_i(\tau) + \frac{G_i}{H_i}u_i(\tau - T_i) \tag{1}$$

where $p_i(\tau)$ is the actual power output of DER i at time τ; $p_i(\tau)$ is equal to $\tilde{p}_i(\tau) + \Delta p_i(\tau)$, in which $\tilde{p}_i(\tau)$ and $\Delta p_i(\tau)$are the intraday power output plan and incremental power output adjustment of DER i; $u_i(\tau - T_i)$ is the reference power output to DER i with time delay T_i; G_i is the gain coefficient that represents the tracking accuracy of the DER i; H_i is the inertia coefficient of DER i.

Note that (1) only presents a rough model to depict the DER dynamic performance, the exact dynamic model of DERs in the practical environment is much more complex and difficult to calibrate. However, the simplified mathematical model of DER can help to formulate an offline simulator to approximate the practical environment and accumulate experiences for the DRL approach.

The regulation ranges of DERs are jointly determined by the current power output, energy state and regulation capacity:

$$\Delta\bar{u}_i(\tau) = \min\left\{\bar{p}_i - \tilde{p}_i(\tau), \left(\tilde{e}_i(\tau) - \underline{e}_i\right)/(h_\rho \Delta t), \bar{r}_i\right\} \tag{2}$$

$$\Delta\underline{u}_i(\tau) = \max\left\{\underline{p}_i - \tilde{p}_i(\tau), (\tilde{e}_i(\tau) - \bar{e}_i)/(h_\rho \Delta t), -\bar{r}_i\right\} \tag{3}$$

where $\Delta\bar{u}_i(\tau)$ and $\Delta\underline{u}_i(\tau)$ are the upper bound and lower bound of the regulation command to DER i; $\tilde{p}_i(\tau)$ is the power output of intraday dispatch plan of DER i; $\underline{p}_i$ and $\bar{p}_i$ are the lower and upper limit of power output of DER i; $\underline{e}_i$ and $\bar{e}_i$ are lower and upper limit of the energy capacity; $\tilde{p}_i(\tau)$ and $\tilde{e}_i(\tau)$ are the active power output and residual energy state of DER i from the intraday dispatch plan; $\bar{r}_i$ is the maximum regulation range offered by DER i; Δt is the scheduling interval of the regulation request; In accordance with [13], [23], to offset the energy deviation of DERs caused by providing frequency regulation service, the concept of the energy reserve ratio (h_ρ) is employed in this study, which enables the DER owners to keep adequate residual energy margin to provide ancillary services continuously for at least h_ρ hour.

The residual energy states of the DERs should be updated after each control cycle as follows:

$$e_i(\tau) = \xi_i e_i(\tau - 1) - \begin{cases} \Delta t \cdot \eta_i^{\text{in}} \cdot p_i(\tau), & p_i(\tau) < 0 \\ \Delta t \cdot p_i(\tau)/\eta_i^{\text{out}}, & p_i(\tau) \geq 0 \end{cases} \tag{4}$$

where $e_i(\tau)$ and $p_i(\tau)$ are the residual energy and power output of DER i during control period τ; η_i^{in} and η_i^{out} are the input and output conversion efficiency; ξ_i represents the energy retention rate.

2) Regulation Cost: The regulation costs of individual DER include the incremental operational cost and regulation mileage cost. The steady-state operation cost function of DERs can be modelled as a quadratic function [26].

$$f_i^{\text{ope}}(p_i(\tau)) = a_{2,i}(p_i(\tau))^2 + a_{1,i}p_i(\tau) + a_{0,i} \tag{5}$$

where $a_{2,i}$, $a_{1,i}$, and $a_{0,i}$ are the steady-state operational cost coefficients of DER i.

Due to that the active power output of the DER is adjusted based on the intraday dispatch plan, the incremental operational cost is introduced to depict the variation of DERs' steady-state operating cost. The incremental operational cost coefficient corresponding to the power output variation can be approximated as follows:

$$c_i(\tau) = \left.\frac{\partial f_i^{\text{ope}}(p_i(\tau))}{\partial p_i(\tau)}\right|_{p_i(\tau) = \tilde{p}_i(\tau)} = 2a_{2,i}\tilde{p}_i(\tau) + a_{1,i} \tag{6}$$

where $c_i(\tau)$ is the incremental operational cost coefficients of DER i.

The frequent regulation request may reduce the service life of DERs [13]. Thus, the regulation mileage cost can represent the friction loss and degradation issue of DERs, which is already considered in many prevailing ancillary service markets [27]. The regulation mileage cost of two adjacent time slots is modelled as follows:

$$f_i^{\text{mil}}(\tau) = b_i|p_i(\tau) - p_i(\tau - 1)| \tag{7}$$

where b_i is the mileage cost coefficient of DER i.

III. Two-Stage DRL Approach for VPP Disaggregation

Managing the DER aggregators in a VPP is a challenging task during the real-time operation since the dynamic parameters of the aggregators are inaccurate and hard to be calibrated. In addition, the frequency regulation request issued by the system operator changes continuously and frequently (typically every few seconds). Thus, during real-time operation, the VPP operator needs to efficiently compute the regulation request disaggregation plans with inaccurate environmental parameters, which is challenging for the traditional model-based controllers or some complex optimization algorithms. The above challenges motivate the DRL approach to enable the VPP to provide frequency regulation service and realize fast decomposition of the regulation request to DER aggregators.

Furthermore, during the start-up process of the training stage, the sampled data is insufficient and the VPP operator has little knowledge about the environment. The DRL algorithm needs to generate many random actions to explore the environment, which would lead to expensive operational costs and significant tracking errors. In light of this, a two-stage DRL approach is proposed in this study to improve the algorithm performance of the online start-up process.

A. Formulation of the Offline Simulator

During the start-up process of the DRL training stage, the sampled data is insufficient, and the agent has little knowledge of the environment. The DRL agent generates a large amount of random data to explore the environment, which would lead to high economic cost and tracking error. Hence, it is necessary to conduct an offline training process to accumulate prior knowledge and data before applying to online implementation. In light of this, an offline simulator is formulated to imitate the practical environment in the real world. Based on the offline simulator, the DRL agent can be trained offline first and then transferred to the online implementation and updated continuously. Therefore, in this subsection, an offline simulator is formulated to approximate the practical environment of the VPP during the offline training stage.

The number of DERs in the power system is huge, which reduces the training efficiency significantly and causes convergence issues for the DRL algorithm. It is difficult to find an effective policy to achieve the optimal management for the DRL in high-dimensional tasks, which motivates this work to manage the DERs in the same aggregator uniformly. By controlling aggregators rather than individual DERs, the dimensions of the state/action space are decreased considerably.

According to the DER model presented in (1)-(4), the dynamic parameters of DERs include χ_i=$\{G_i, H_i, T_i, \eta_i^{\mathrm{in}}, \eta_i^{\mathrm{out}}, \xi_i\}$ and ϑ_i=$\{\tilde{p}_i(\tau), \bar{p}_i, \underline{p}_i, \tilde{e}_i(\tau), \bar{e}_i, \underline{e}_i, \bar{r}_i\}$. To facilitate the management of numerous devices, DERs can be classified into different aggregators based on the constant parameters among χ_i and ϑ_i.

The approximation parameters of aggregators can be obtained by computing the geometric centre of χ_i and summation of ϑ_i in the same aggregator.

$$\chi_K = \sum_{i\in\Theta_K^{\mathrm{AGG}}} (w_i \chi_i) \tag{8}$$

$$\vartheta_K = \sum_{i\in\Theta_K^{\mathrm{AGG}}} (\vartheta_i) \tag{9}$$

where Θ_K^{AGG} represents the set of DERs in aggregator K; χ_K and ϑ_K are the approximation parameters of aggregator K, χ_K=$\{G_K,H_K,T_K,\eta_K^{\mathrm{in}},\eta_K^{\mathrm{out}},\xi_K\}$, ϑ_K=$\{\tilde{P}_K(\tau),\bar{P}_K,\underline{P}_K,\tilde{E}_K(\tau),\bar{E}_K,\underline{E}_K,\bar{R}_K\}$; w_i is the weight factor of DER i satisfying $\sum_{i\in\Theta_K^{\mathrm{AGG}}} w_i = 1$. In accordance with [28]–[30], w_i can be determined by the rated active power capacity of DER i or the scaling factor defined in [23].

The approximation of the dynamic model of the DER aggregators is presented as follows:

$$\dot{P}_K(\tau) = -\frac{1}{H_K}P_K(\tau) + \frac{G_K}{H_K}\big(\Delta U_K(\tau - T_K) + \tilde{P}_K(\tau)\big) \tag{10}$$

$$E_K(\tau) = \xi_K E_K(\tau-1) - \begin{cases} \Delta t \cdot \eta_K^{\mathrm{in}} \cdot P_K(\tau), & P_K(\tau) < 0 \\ \Delta t \cdot P_K(\tau)/\eta_K^{\mathrm{out}}, & P_K(\tau) \geq 0 \end{cases} \tag{11}$$

$$\Delta\underline{U}_K \leq \Delta U_K(\tau) \leq \Delta\bar{U}_K \tag{12}$$

$$\Delta\bar{U}_K(\tau) = \min\big\{\bar{P}_K - \tilde{P}_K(\tau), \big(\tilde{E}_K(\tau) - \underline{E}_K\big)/\big(h_\rho \Delta t\big), \bar{R}_K\big\} \tag{13}$$

$$\Delta\underline{U}_K(\tau) = \max\big\{\underline{P}_K - \tilde{P}_K(\tau), \big(\tilde{E}_K(\tau) - \bar{E}_K\big)/\big(h_\rho \Delta t\big), -\bar{R}_K\big\} \tag{14}$$

where $P_K(\tau)$ and $E_K(\tau)$ are the power output and residual energy of aggregator K at time τ; $\Delta U_K(\tau)$ is the regulation command on the basis of intraday dispatch plan to aggregator K at time τ; $\Delta\bar{U}_K$ and $\Delta\underline{U}_K$ are the upper bound and lower bound of the regulation command to aggregator K; $\tilde{P}_K(\tau)$ and $\tilde{E}_K(\tau)$ are the intraday dispatch plans of active power and residual energy of aggregator K at time τ; $\underline{P}_K$, $\bar{P}_K$, $\underline{E}_K$, and $\bar{E}_K$ are the lower and upper limit of power output and energy state of aggregator K; $\bar{R}_K$ is maximum regulation range of aggregator K.

According to regulation cost of individual DERs presented in (6) and (7), the approximation of the regulation cost functions of the aggregators are formulated as follows:

$$F_K^{\mathrm{ope}}(\tau) = C_K(\tau)\Delta P_K(\tau) \tag{15}$$

$$F_K^{\mathrm{mil}}(\tau) = B_K|P_K(\tau) - P_K(\tau-1)| \tag{16}$$

where $\Delta P_K(\tau) = P_K(\tau) - \tilde{P}_K(\tau)$ is the incremental power output adjustment of aggregator K; $F_K^{\mathrm{ope}}(\tau)$ and $F_K^{\mathrm{mil}}(\tau)$ are the steady operational cost and the regulation mileage cost of aggregator K; $C_K(\tau)$ and B_K are the approximation of incremental operational cost coefficient and mileage cost coefficient of aggregator K,$C_K(\tau) = \sum_{i\in\Theta_K^{\mathrm{AGG}}}(w_i c_i(\tau))$,$B_K = \sum_{i\in\Theta_K^{\mathrm{AGG}}}(w_i b_i)$.

B. Online Implementation

After experience accumulation and policy training in the offline simulator, the offline policy is transferred to online-stage and updated continuously in the practical environment. During the initial stage of online implementation, the tracking accuracy and operational economy of the practical VPP environment can be enhanced considerably, since DRL agent have accumulated adequate knowledge during the offline-stage. Note that, the definitions of the episode, state space, action space, reward, and variables of the online MDP problem are the same as that in offline simulation stage. Due to page limitation, we only list the differences between the two stages.

Different from the offline-stage, the state variables and rewards are obtained from the measurement and observation of the practical environment. Therefore, the following variables, parameters, and operational costs of the aggregators should be updated according to actual telemetry of individual DERs.

$$P_K(\tau) = \sum_{i\in\Theta_K^{\mathrm{AGG}}} (p_i(\tau)), E_K(\tau) = \sum_{i\in\Theta_K^{\mathrm{AGG}}} (e_i(\tau)) \tag{17a}$$

$$\Delta\bar{U}_K(\tau) = \sum_{i\in\Theta_K^{\mathrm{AGG}}} (\Delta\bar{u}_i(\tau)), \Delta\underline{U}_K(\tau) = \sum_{i\in\Theta_K^{\mathrm{AGG}}} \left(\Delta\underline{u}_i(\tau)\right) \tag{17b}$$

$$F_K^{\mathrm{ope}}(\tau) = \sum_{i\in\Theta_K^{\mathrm{AGG}}} \left[f_i^{\mathrm{ope}}(p_i(\tau)) - f_i^{\mathrm{ope}}(\tilde{p}_i(\tau))\right] \tag{17c}$$

$$F_K^{\mathrm{mil}}(\tau) = \sum_{i\in\Theta_K^{\mathrm{AGG}}} \left(f_i^{\mathrm{mil}}(\tau)\right). \tag{17d}$$

Moreover, the regulation command should be decomposed to all the individual DERs during online implementation process. The decomposition principle from aggregators to individual DERs is defined as follows:

$$\Delta u_i(\tau) = \begin{cases} \frac{\Delta\bar{u}_i}{\Delta\bar{U}_K}\left(U_K(\tau) - \tilde{P}_K(\tau)\right), U_K(\tau) \geq \tilde{P}_K(\tau) \\ \frac{\Delta\underline{u}_i}{\Delta\underline{U}_K}\left(U_K(\tau) - \tilde{P}_K(\tau)\right), U_K(\tau) < \tilde{P}_K(\tau). \end{cases} \tag{18}$$

Although the further disaggregation method in (18) cannot guarantee the optimality of the solution, the following two reasons indicate the necessity of using this method: i). Combining with (2)-(3), (18) can ensure the solution feasibility and realize the fast decomposition of the regulation commands during real-time operation. ii). This method enables all the adjustable DERs to share the superior regulation request and provide regulation services simultaneously. Hence, the residual energy deviations of all the DERs are small and change synchronously, which helps to enhance the consistency between the real-time energy states and the intraday dispatch plans and promote the long-term economical operation of the VPP.

C. Markov Decision Process and Justification of Using DRL

The VPP operator is considered as the intelligent DRL agent and the physical system is considered as the environment. The VPP operator needs to make sequential decisions in an uncertain environment and maximize the cumulative reward. Mathematically, the decision-making problem should be modelled as the *Markov Decision Processes* (MDP). The MDP is defined by the tuple $M(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$ with the following basic assumptions and conditions [31]-[33].

i). $\mathcal{S}$ is a finite state space.

ii). $\mathcal{A}$ is a finite action space.

iii). $r(\boldsymbol{s}, \boldsymbol{a}) : \mathcal{S} \times \mathcal{A} \rightarrow \mathrm{R}$ is the bounded reward.

iv). γ is the discounting factor that penalizes future rewards.

v). $\mathcal{P}$ is the transition kernel with transition probability$(\cdot|\boldsymbol{s}, \boldsymbol{a}) : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$, through which the current state-action pair $(\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$ is mapped to a distribution on the next state $\boldsymbol{s}' \in \mathcal{S}$; hence, the next transition should only depend on the current states and the current actions.

Note that the basic assumptions and conditions of MDP listed in i)-iv) are valid for the problem being addressed. The details are explained as follows: According to the definitions of states, actions, and reward in Section III-C, the dimensions of the state space and action space are finite and the reward is bounded, which satisfies the assumptions and conditions listed in i)-iii); The discounting factor (γ) listed in iv) is set less than 1.0 and incorporated into the objective of the DRL algorithm in (26), which ensures that the future reward is discounted over time. In addition, according to the dynamic characteristic presented in (1)-(4) and (10)-(14), the next states (P_K, E_K, $\tilde{P}_K$, $\tilde{E}_K$, $\Delta\bar{U}_K$, $\Delta\underline{U}_K$) only depend on the current states and actions. Furthermore, in accordance with [14]–[18], under the mild assumptions that the uncertain parameters (including renewable energy generation, load demand, and regulation command) in the environment obey certain probability distributions, the transition probability of the next state (δ^{VPP}) is also determined by the current states and actions according to (24). Therefore, assumption v) is also satisfied since the next transition only depends on the current states and actions. In light of this, it is justified to utilize the DRL approach to solve the regulation service disaggregation problem in a VPP.

D. Definition of Action Space, State Space, and Reward

The definitions of the state space, action space, and reward function of the offline DRL agent are presented as follows. The continuous control problems for both offline and online-stage are considered. Thus, the episode is infinite.

i). *Action ($\boldsymbol{a}_\tau \in \mathcal{A}$):* The action space includes the regulation ratio $\boldsymbol{a}_\tau = \left[\boldsymbol{a}_{K,\tau}\right]$ to all the DER aggregators. The regulation ratio is transformed into regulation commands as follows:

$$\Delta U_K(\tau) = \Delta\underline{U}_K(\tau) + \boldsymbol{a}_{K,\tau}\left(\Delta\bar{U}_K(\tau) - \Delta\underline{U}_K(\tau)\right) \tag{19}$$

where $\boldsymbol{a}_{K,\tau} \in [0, 1]$ is the action variable corresponding to the regulation ratio to aggregator K at time τ.

ii). *State ($\boldsymbol{s}_\tau \in \mathcal{S}$):* The state space of the MDP problem includes $\boldsymbol{s}_\tau = [\boldsymbol{s}_{K,\tau}]$, where $\boldsymbol{s}_{K,\tau} = \{\delta^{\mathrm{VPP}}(\tau), P_K(\tau), E_K(\tau), \tilde{P}_K(\tau), \tilde{E}_K(\tau), \Delta\bar{U}_K(\tau), \Delta\underline{U}_K(\tau)\}$; δ^{VPP} is the tracking error between the reference regulation request issued by the system operator and the aggregate power output of the VPP.

iii). *Reward ($r_\tau \in \mathbb{R}$):* The reward is defined as follows:

$$r_\tau = \omega^{\mathrm{RM}}R^{\mathrm{RM}}(\tau) - \sum_{K=1}^{N^{\mathrm{AGG}}}\left[\omega^{\mathrm{cost}}\left(F_K^{\mathrm{ope}}(\tau) + F_K^{\mathrm{mil}}(\tau)\right) + \omega^{\mathrm{pnl}}F_K^{\mathrm{pnl}}(\tau)\right] \tag{20}$$

where $R^{\mathrm{RM}}(\tau)$ is the revenue from providing regulation service; $F_K^{\mathrm{pnl}}(\tau)$ is the penalty for the deviation between regulation trajectory and intraday dispatch plan of aggregator K; N^{AGG} is the number of aggregators; ω^{RM}, ω^{cost}, and ω^{pnl} are the weight coefficients corresponding to different objectives.

1) Revenue From Providing Regulation Service: Based on the prevailing performance-based regulation service markets, the VPP revenue from providing regulation service is simultaneously determined by the regulation mileage deployment and tracking accuracy. Inspired by the market settlement mechanism of CAISO, the regulation mileage settlement revenue is adopted to quantify the reward from providing regulation service [3], [27].

$$R^{\mathrm{RM}}(\tau) = \rho^{\mathrm{VPP}}(\tau)\pi^{\mathrm{MP}}(\tau)M^{\mathrm{VPP}}(\tau) \tag{21}$$

where $\pi^{\mathrm{MP}}(\tau)$ is the mileage price in the regulation service market at time τ; $M^{\mathrm{VPP}}(\tau)$ is the instruct regulation mileage between two adjacent time slots issued by the system operator to the VPP agent at time τ; $\rho^{\mathrm{VPP}}(\tau)$ is the performance accuracy adjustment coefficient at time τ.

$$M^{\mathrm{VPP}}(\tau) = \left|U^{\mathrm{VPP}}(\tau) - U^{\mathrm{VPP}}(\tau - 1)\right| \tag{22}$$

$$\rho^{\mathrm{VPP}}(\tau) = \frac{U^{\mathrm{VPP}}(\tau) - \left|\delta^{\mathrm{VPP}}(\tau)\right|}{U^{\mathrm{VPP}}(\tau)} \tag{23}$$

$$\delta^{\mathrm{VPP}}(\tau) = U^{\mathrm{VPP}}(\tau) - \sum_{K=1}^{N^{\mathrm{unit}}} P_K(\tau) - \sum_{n=1}^{N^{\mathrm{RES}}} P_n^{\mathrm{RES}}(\tau) + P^{\mathrm{L}}(\tau) \tag{24}$$

where $U^{\mathrm{VPP}}(\tau)$ is the reference regulation request at time τ; N^{RES} and N^{unit} are the number of renewable energy sources and controllable units; $P_n^{\mathrm{RES}}(\tau)$ is the active power output of the n^{th} renewable energy source at time τ; $P^{\mathrm{L}}(\tau)$ is the load demand in the VPP at time τ.

Note that the proposed model only needs to make minor adjustments when extending to many other performance-based regulation service markets, such as the PJM, MISO, and NYISO in North America, Guangdong provincial power market in China, and several European TSOs [2], [3], [34], since such settlement mechanism based on the accuracy performance and regulation mileage is generally applicable.

The penalty imposed on the deviation between regulation commands and intraday dispatch plan is defined as follows:

$$F_K^{\mathrm{pnl}}(\tau) = \omega^P\big(P_K(\tau) - \tilde{P}_K(\tau)\big)^2 + \omega^E\big(E_K(\tau) - \tilde{E}_K(\tau)\big)^2 \tag{25}$$

where $F_K^{\mathrm{pnl}}(\tau)$ is the penalty for power and energy deviation between the regulation trajectory and intraday dispatch plan; ω^P and ω^E are the weight coefficients for the power deviation and energy deviation.

IV. Methodology

The DRL agent aims to find the control policy to maximize the expected cumulative discounted reward. The actor-critic framework is a widely used architecture in the DRL methods, in which the actor network updates the target policy based on gradient ascent and the critic network estimates the expected Q-value. The existing actor-critic DRL algorithms can be divided into two categories: on-policy (such as asynchronous advantage actor-critic and proximal policy optimization) and off-policy (deep deterministic policy gradient and soft actor-critic). Compared with the on-policy methods, the off-policy methods are more appealing for power systems since any policy or historical operating scenario could generate the training data. In addition, compared with other off-policy methods such as deterministic policy gradient algorithm, soft actor-critic (SAC) is a sample-efficient and prominent method to learn a stochastic policy, which can enhance the exploration of the DRL agent with a faster convergence speed [19], [20].

A. The Soft Actor-Critic Algorithm

The maximum entropy objective is adopted for SAC, in which the common objective is augmented with an entropy term to encourage exploration.

$$\pi^* = \arg\max_{\pi} V^{\pi}(\boldsymbol{s}) = \mathop{\mathbb{E}}_{\boldsymbol{a}_\tau \sim \pi(\cdot|\boldsymbol{s}_\tau)} \sum_{\tau=0}^{\infty} \gamma^{\tau}\big[r_\tau(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau) - \alpha \log \pi(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau)\big] \tag{26}$$

where $\pi(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau)$ is the policy for the agent that maps from the state space $\mathcal{S}$ to a distribution on the action space $\mathcal{A}$; α is the temperature parameter that determines the weight of the entropy versus the reward; $\gamma \in (0, 1)$ is the discounting factor that penalizes the future rewards.

The SAC algorithm has two neural networks: the Q network and the policy network. The Q network parameterized by $\theta^{(n)}$ is employed to approximate the soft Q-function $Q_{\theta^{(n)}}(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau)$ of state-action pairs $(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau)$, and the policy network parameterized by ϕ takes in the state and outputs the mean and deviation of actions' probability distribution.

Considering automating entropy adjustment, the gradient for α is computed by minimizing the following objective according to [20].

$$\min_{\alpha} J(\alpha) = \mathbb{E}_{\boldsymbol{a}_\tau \sim \pi_\phi}\big[-\alpha \log \pi_\phi(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau) - \alpha\bar{\mathcal{H}}\big] \tag{27}$$

where $\bar{\mathcal{H}}$ is the desired minimum expected entropy.

The policy network parameters (ϕ) can be trained to minimize the expected Kullback-Leibler divergence as follows:

$$\begin{aligned}\min_{\phi} J_\pi(\phi) \\ = \mathbb{E}_{\boldsymbol{s}_\tau \sim \mathcal{B}, \varsigma_\tau \sim \mathcal{N}}\big[\alpha \log\big(\pi_\phi\big(f_\phi(\varsigma_\tau; \boldsymbol{s}_\tau)|\boldsymbol{s}_\tau\big)\big) - Q_\theta\big(\boldsymbol{s}_\tau, f_\phi(\varsigma_\tau; \boldsymbol{s}_\tau)\big)\big]\end{aligned} \tag{28}$$

where $\mathcal{B}$ represents the distribution of sampled states and actions or the replay memory buffer; $\boldsymbol{a}_\tau = f_\phi(\varsigma_\tau; \boldsymbol{s}_\tau)$ represents that the policy is reparameterized using the neural network transformation; ς_τ is the input noise vector sampled from spherical Gaussian distribution.

The Q network parameters ($\theta^{(n)}$) should be learned by minimizing the soft Bellman residual.

$$\begin{aligned}\min_{\theta^{(n)}} J_Q\Big(\theta^{(n)}\Big) = \mathbb{E}_{(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau) \sim \mathcal{B}}\Bigg[\frac{1}{2}\big(Q_{\theta^{(n)}}(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau) - (r_\tau(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau) \\ + \gamma \mathbb{E}_{\boldsymbol{s}_{\tau+1} \sim \mathcal{P}}\big[Q_{\bar{\theta}}(\boldsymbol{s}_{\tau+1}, \boldsymbol{a}_{\tau+1}) - \alpha \log\big(\pi_\phi(\boldsymbol{a}_{\tau+1}|\boldsymbol{s}_{\tau+1})\big)\big]\big)\big)^2\Bigg].\end{aligned} \tag{29}$$

where $\bar{\theta}$ represents the target Q network that is equal to the exponentially moving average of the Q network.

Furthermore, for the SAC algorithm, the policy π_ϕ is a complicated function related to the temperature parameter α, and the neural network transformations are adopted to parameterize the policy and soft Q-function. Therefore, the minimization problems in (27)-(29) are very complex. To avoid directly computing the exact gradients, the state-of-the-art literature has proposed a series of math tricks to find the approximate gradients of problems (27)-(29) [19], [20]. The details are presented as follows:

$$\hat{\nabla} J(\alpha) = -\log \pi_\phi(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau) - \bar{\mathcal{H}} \tag{30}$$

$$\begin{aligned}\hat{\nabla} J_\pi(\phi) = \nabla_\phi \alpha \log\big(\pi_\phi(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau)\big) \\ + \big(\nabla_{\boldsymbol{a}_\tau} \alpha \log\big(\pi_\phi(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau)\big) - \nabla_{\boldsymbol{a}_\tau} Q_\theta(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau)\big)\nabla_\phi f_\phi(\varsigma_\tau; \boldsymbol{s}_\tau)\end{aligned} \tag{31}$$

$$\begin{aligned}\hat{\nabla} J_Q(\theta^{(n)}) = \nabla_{\theta^{(n)}} Q_{\theta^{(n)}}(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau)\big(Q_{\theta^{(n)}}(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau) - r_\tau(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau) \\ -\gamma\big(Q_{\bar{\theta}}(\boldsymbol{s}_{\tau+1}, \boldsymbol{a}_{\tau+1}) - \alpha \log\big(\pi_\phi(\boldsymbol{a}_{\tau+1}|\boldsymbol{s}_{\tau+1})\big)\big)\big).\end{aligned} \tag{32}$$

Note that, in accordance with [20], [35], to mitigate the overestimation bias in the iteration that degrades the algorithm performance, two soft Q-functions parameterized by $\theta^{(n)}|_{n=1,2} = \{\theta^{(1)}, \theta^{(2)}\}$ are trained independently to optimize $J_Q(\theta^{(n)})$. Afterwards, the minimum-Q is used in the update of the policy network (28) and the Q network (29). The minimum of the two soft Q-functions is defined as follows.

$$Q_\theta = \min_{n\in\{1,2\}} \left\{Q_{\theta^{(n)}}\right\} \tag{33a}$$

$$Q_{\bar{\theta}} = \min_{n\in\{1,2\}} \left\{Q_{\bar{\theta}^{(n)}}\right\}. \tag{33b}$$

B. The Proposed SAM-SAC Algorithm

There always exist inevitable discrepancies between the offline and online environments. To improve the adaptability and robustness of the SAC algorithm, a sharpness-aware minimization based soft actor-critic (SAM-SAC) algorithm is *firstly* proposed in this study, which aims to alleviate the adverse impact caused by discrepancies.

Sharpness-aware minimization (SAM) minimizes both the loss value and the loss sharpness simultaneously [21], [22]. This method seeks the objectives within a neighbourhood that have a uniformly low loss. SAM shows decent performance in enhancing the neural networks' generalization, accuracy, and robustness. In light of this, SAM is employed in the Q network in the SAC algorithm. According to [22], the loss function of the Q network in each iteration satisfies:

$$J_Q(\theta_i) \le \max_{\left\|T_\theta^{-1}\varepsilon\right\|_2 \le \rho} J_Q(\theta_i + \varepsilon) + h\left(\frac{\|\theta_i\|_2^2}{\rho^2}\right) \tag{34}$$

where $J_Q(\theta + \varepsilon)$ is the training loss representing the model generalization ability; $h : \mathrm{R}^+ \to \mathrm{R}^+$ is a strictly increasing function; ρ is a hyper-parameter impacting the concerned neighbourhood; T_θ is the element-wise normalization operator that satisfies the scale-invariant property [22].

Therefore, the original loss function of the Q network in (29) can be redefined by the following adaptive sharpness-aware minimization problem:

$$\min_\theta \max_{\left\|T_\theta^{-1}\varepsilon\right\|_2 \le \rho} J_Q(\theta + \varepsilon) + \frac{\lambda}{2}\|\theta\|_2^2 \tag{35}$$

where λ is the weight decaying coefficient of ℓ^2 regularization.

Compared with the traditional SAC, the major innovation of the proposed SAM-SAC algorithm is to advance the training loss of the Q network from (29) to (35). Q network aims to estimate the environmental reward by training a set of parameters represented by θ. If considering ε as the environmental parameter disturbance, the bi-level robust optimization model in (35) can enhance the robustness of the training objective against disturbances. Thus, the adaptability of the DRL algorithm to the noises and environmental parameter disturbances can be improved.

The min-max problem (35) can be solved by a two-step iteration. The optimal ε is found in the inner layer first and then θ is computed in the outer layer. The approximation to solve the min-max problem in (35) can refer to the Appendix. The two-step iteration over training steps to find the approximations to the optimal values of ε and θ are expressed as follows:

$$\varepsilon_k = \rho \frac{T_{\theta_k}^2 \hat{\nabla} J_Q(\theta_k)}{\left\|T_{\theta_k}\hat{\nabla} J_Q(\theta_k)\right\|_2} \tag{36}$$

$$\theta_{k+1} = \theta_k - \delta_Q\left(\hat{\nabla} J_Q(\theta_k + \varepsilon_k) + \lambda\theta_k\right) \tag{37}$$

where ε_k and θ_k are the approximation of the optimal values of the inner layer problem and outer layer problem at training step k, respectively; δ_Q is the learning rate of the Q network.

The SAM method originates from [21] and [22], which can improve the model robustness and generalization performance of the neural networks. The purpose of the Q network is to predict the future reward by fitting the environment of the DRL algorithm. The SAM method can help the Q network to provide higher robustness and adaptability to environmental noise and parameter disturbance. Hence, the technical advantage of the SAM method is quite in accordance with the functions and purposes of the Q network. Nevertheless, the policy network aims to update the control policy by minimizing the Kullback-Leibler divergence, which does not need to fit or predict the environment of the DRL algorithm [19], [20]. Therefore, the function of the policy network does not match with the technical advantage of the SAM method. Furthermore, the two-step iteration in (36)-(37) to solve the SAM problem will introduce extra computation burden and reduce the training efficiency. Therefore, in the proposed algorithm, the SAM method is only used for the Q network.

Some advanced techniques can be easily incorporated into the proposed SAM-SAC algorithm. Moreover, since the DRL agent sequentially interacts with the environment and samples from the system observation, the adjacent sampled states are not identically independent. Therefore, an experience reply buffer $\mathcal{B}$ to store the previous experiences $(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau, r_\tau, \boldsymbol{s}_{\tau+1})$ is utilized in the proposed SAM-SAC algorithm [15], and a batch of experiences is sampled uniformly from the buffer to update the neural networks at every environment step. To accelerate the training process, prioritized experience replay (PER) [36] and emphasizing recent experience without forgetting the past (EREFP) [37] technologies are employed to enhance the sampling efficiency and replay important transitions.

The pseudocode of the proposed SAM-SAC algorithm is presented in Algorithm 1.

Note that, the proposed SAM-SAC algorithm is also applicable to many other scenarios, such as frequency control [14], economic dispatch [16], and voltage regulation [18]. To the authors' best knowledge, few works study the VPP regulation service disaggregation problem using the DRL approach. Therefore, this study could fill the gap in the existing literature and expand the application scope of the DRL method in power system operations.

C. Two-Stage Implementation of the Proposed DRL Approach

Following Section III, to alleviate the expensive operational cost and significant tracking error of the online start-up process

Algorithm 1 : SAM-SAC (Pseudocode)

```
Input: $\theta^{(1)}$, $\theta^{(2)}$, $\bar{\theta}^{(1)}$, $\bar{\theta}^{(2)}$-initial parameters of the $Q$ network; $\phi$-initial parameters of the policy network; $\mathcal{B} \leftarrow \emptyset$ initialize an empty replay buffer; $\upsilon$-soft updating factor; $\delta_Q$,$\delta_\pi$, $\delta_\alpha$-learning rate of $Q$ network, policy network, and $\alpha$.
for each iteration do
    for each environment step do
        $\boldsymbol{a}_\tau \sim \pi_\phi(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau)$;
        $\boldsymbol{s}_{\tau+1} \sim \mathcal{P}(\boldsymbol{s}_{\tau+1}|\boldsymbol{s}_\tau, \boldsymbol{a}_\tau)$;
        $\mathcal{B} \leftarrow \mathcal{B} \cup \{(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau, r_\tau, \boldsymbol{s}_{\tau+1})\}$;
    end for
    for each training step do
        Sample a batch of experiences from $\mathcal{B}(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau, r_\tau, \boldsymbol{s}_{\tau+1})$;
        Compute the minimum-$Q$ ($Q_\theta$,$Q_{\bar{\theta}}$) using (33);
        Substitute $Q_\theta$ and $Q_{\bar{\theta}}$ into the update of the policy network (28) and the $Q$ network (29), and compute the gradients of $\hat{\nabla} J(\alpha_k)$, $\hat{\nabla} J_\pi(\phi_k)$, $\hat{\nabla} J_Q(\theta_k^{(n)})$ based on (30)-(32);
        $\varepsilon_k^{(n)} \leftarrow \rho \frac{T^2_{\theta_k^{(n)}} \hat{\nabla} J_Q(\theta_k^{(n)})}{\|T_{\theta_k^{(n)}} \hat{\nabla} J_Q(\theta_k^{(n)})\|_2}$ for $n \in \{1, 2\}$;
        $\theta_{k+1}^{(n)} \leftarrow \theta_k^{(n)} - \delta_Q(\hat{\nabla} J_Q(\theta_k^{(n)} + \varepsilon_k^{(n)}) + \lambda\theta_k^{(n)})$ for $n \in \{1, 2\}$;
        $\phi_{k+1} \leftarrow \phi_k - \delta_\pi \hat{\nabla} J_\pi(\phi_k)$;
        $\alpha_{k+1} \leftarrow \alpha_k - \delta_\alpha \hat{\nabla} J(\alpha_k)$;
        $\bar{\theta}_{k+1}^{(n)} \leftarrow \upsilon\theta_{k+1}^{(n)} + (1-\upsilon)\bar{\theta}_k^{(n)}$ for $n \in \{1, 2\}$;
    end for
end for
Output $\theta_k^{(1)}$,$\theta_k^{(2)}$, $\phi_k$.
```

caused by algorithm exploration, the proposed SAM-SAC algorithm is trained based on a two-stage DRL framework, as presented in Fig. 2. The algorithm is implemented in the offline simulator to accumulate prior experiences and train an initial policy for the online start-up process first, and then the obtained control policy is transferred to online implementation and updated continuously.

In the offline training stage, a virtual environment in the offline simulator is formulated to approximate the practical environment. The approximation methods of DER aggregators' characteristics are presented in (10)-(16), (25). The reward and state variables originate from observing the virtual environment in the offline simulator. The action variables are converted to regulation commands issued to controllable units and implemented in the offline simulator. The proposed DRL approach is implemented based on the actor-critic framework to train the VPP disaggregation policy. The algorithm has two categories of neural networks, namely, the Q network and the policy network. The Q network takes in the state-action pairs $(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau)$ and outputs the soft Q-function to approximate the reward from the environment. The policy network takes in the state variables and outputs the mean and deviation of actions' distribution. The loss functions of the policy network and Q network are listed in (28) and (35). The parameters of these neural networks are updated using the proposed SAM-SAC algorithm presented in Algorithm 1.

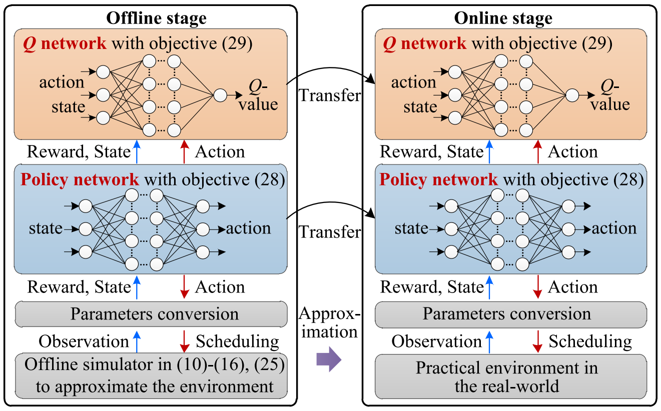


Fig. 2. Two-stage implementation approach of the proposed DRL approach.

After the data accumulation and policy training in the offline simulator, the parameters of the neural networks is transferred to the online-stage and updated continuously in the practical environment. The DRL algorithm of the online-stage is the same as that of the offline-stage. The major difference is that the action variables are implemented in the practical environment, and the reward/state variables are obtained from the measurement of the real world. The performance of the online start-up process can be improved based on the two-stage implementation approach since the initial disaggregation policy for online implementation has accumulated prior knowledge in the offline training stage.

Moreover, the applicability and limitations of the proposed approach are further discussed and justified as follows.

i). The dynamic characteristic of the DERs depicted in (1) is essentially a rough model, which has been used to approximate the real model of DERs in the existing literature, as in [24], [25]. Although calibrating the parameters of DERs' dynamic characteristics is not the main focus of this study, it has been profoundly investigated in the existing literature. Some advanced parameter estimation methods, e.g., the closed-loop identification method [30], mode decomposition [38] and recursive approach [39], can be adopted to enhance the approximation accuracy of the dynamic parameters. Likewise, in accordance with [28]–[30], a similar formulation is introduced in (10) to approximate the aggregated dynamic model of DERs. Note that, despite the inevitable discrepancies between the offline simulation and the practical environment, the above offline simulator is still helpful to accumulate data/experience and train an initial control policy with a decent start-up process for the online implementation.

ii). The DRL algorithm can directly compute the decisions according to the sampling and measurement of the current state variables in the VPP. Thus, although the mathematical model of the DER aggregators is formulated in the offline simulator, the proposed approach does not rely on the complete information of the environment.

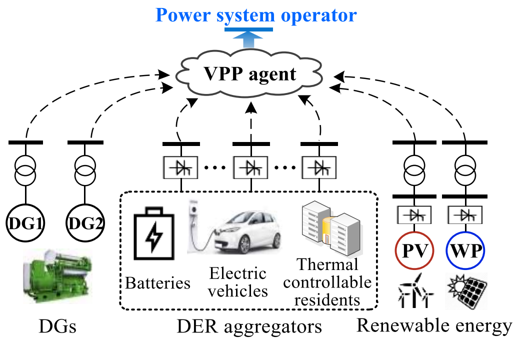


Fig. 3. Components of the virtual power plant.

iii). The regulation service disaggregation problem of the aggregators with time-varying types and number of DERs is out of the scope of this study, which can be a subject of the future work. The reason is explained as follows: when the proposed approach is implemented to the real-world environment, the DER aggregators need to monitor and observe the current states (including energy, power and regulation range) of the internal DER devices. Therefore, the type and number of the maximum available DER devices in the same aggregators should be predetermined. Despite such limitations, the proposed approach is still promising the scenarios that the ownership between the individual DER owners and aggregators is time-invariant, as in [8], [10]–[12].

V. Case Study

To investigate the performance and effectiveness of the proposed approach, this section presents the simulation results of a VPP consisting of renewable energy generation, conventional generators, and various types of DERs, as in Fig. 3.

A. Simulation Environment Setup

The algorithm is programmed in Python using PyCharm, while the multilayer neural networks utilized in the DRL approach are formulated using PyTorch. All the tests were conducted on a computer with an Intel i7-3610QM CPU and 16 GB of RAM. The VPP contains a PV station (10MW) and a wind farm (10MW). The sequential RES generation curves are obtained from the real data from the Jilin provincial power grid in northeastern China. RESs are operated at the maximum power point tracking mode.

Without loss of generality, 3000 DERs (1000 electric vehicle charging loads, 1000 thermal controllable residents, and 1000 energy storage batteries) are included in the test system and classified into 8 aggregators, as in [23]. The dynamic characteristic parameters and operation cost coefficients of the DERs are presented in Table I. Moreover, due to page limitation, other relevant parameters of DERs, including the capacity of the power and energy, energy retention rate, and initial energy state, can refer to the supplemental materials in [40]. The parameters of individual DERs are procured using the Monte-Carlo sampling method according to the probability distribution presented in Table I and [40]. The parameters of distributed generators (DGs) are presented in Table II.

TABLE I
PROBABILITY DISTRIBUTION OF DERS' PARAMETERS

Parameters	Distr.	Mean	Deviation	Min	Max
Inertia coefficient H_i (s)	UD	3.5	2.0	0.2	8.0
Time delay T_i (s)	UD	0.45	0.2	0.2	0.8
Cost coefficient $a_{2,i}$ (\$/MWh^2)	TGD	15.4	5	5	30
Cost coefficient $a_{1,i}$ (\$/MWh)	TGD	0.5	0.25	0.15	1
Cost coefficient $a_{0,i}$ (\$)	TGD	0.05	--	0	0.1
Cost coefficient b_i (\$/MW)	TGD	1.35	0.75	0.3	4.5
Uncontrollable probability (%)	UD	10	2.5	5	20

*TGD: truncated Gaussian distribution. UD: uniform distribution.

TABLE II
PARAMETERS OF CONVENTIONAL GENERATORS

Unit	H_i	T_i	Range (MW)	Cost coefficients			
				$a_{2,i}$	$a_{1,i}$	$a_{0,i}$	b_i
DG1	4.6	0.3	[2,5]	2.8	32.0	1.2	1.2
DG2	12.2	0.65	[2,5]	3.2	29.5	2.6	1.45

TABLE III
ALGORITHM HYPER PARAMETERS

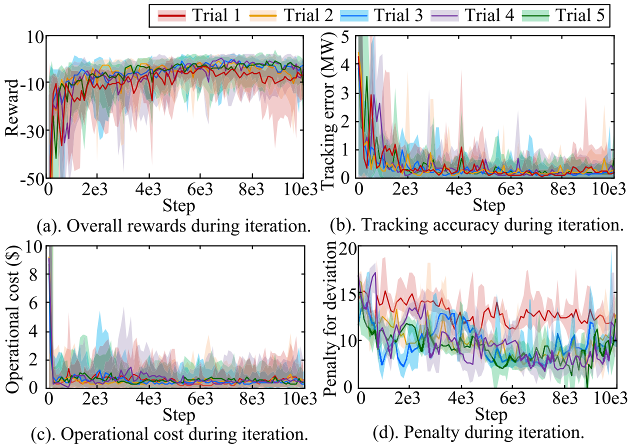
Parameters	Value
Optimizer	Adam
Types of the neural network	Feedforward
Hidden layers number of the neural network	2
Neurons number of each hidden layer	512
Replay buffer size	1e5
Batch size	64
Learning rate	1e-3
Discount factor	0.9
Soft updating factor	1e-2
Moment	0.7
Weight decay	1e-3
Hyper-parameter ρ	5e-3
Energy reserve ratio h_ρ	0.25
ω^{RM} / ω^{cost} / ω^{pnl} / ω^{P} / ω^{E}	10/5/2/0.9/0.1

The hyper parameters of the DRL algorithm are presented in Table III. Note that the uncontrollable probability presented in Table I represents the probability that DERs does not follow the VPP regulation command in the practical environment.

Moreover, to procure adequate data and accumulate enough experience, sequential daily operation scenarios of the VPP, including the intraday dispatch plans issued by the superior market operator and the intraday dispatch commands to the controllable units, are simulated based on historical renewable energy generation curve and market price curve. After that, to imitate the time-varying regulation request issued by the market operator, the training set is generated continuously and recurrently by adding random noises and fluctuations to the intraday dispatch plans using the Monte-Carlo sampling and interpolation methods. The test set is a fixed dataset of 1e4 environment steps.

B. Effectiveness of the Proposed SAM-SAC Algorithm

The reward, absolute value of the tracking error, system operational cost, and penalty for the deviation between the intraday dispatch plan and the VPP actual power output are



(a). Overall rewards during iteration. (b). Tracking accuracy during iteration.

(c). Operational cost during iteration. (d). Penalty during iteration.

Fig. 4. Training process of the proposed SAM-SAC algorithm.

presented in Fig. 4. The traces with the shaded region can reflect the convergence speed and stability of the algorithms.

In Fig. 4, five different trials with different random seeds and training datasets are trained using the proposed SAM-SAC algorithm separately on the offline simulator, and each performing one evaluation rollout every 100 environment steps. The solid line corresponds to the average step-wise returns (including reward, tracking error, cost, operational cost, and penalty for deviation) of every 100 environment steps (100 batches) over each trial. The minimum and maximum boundaries of the shaded region correspond to the minimum and maximum step-wise returns of every 100 environment steps over each trial.

Simulation results in Fig. 4 reveal that the overall reward, absolute value of the tracking error, system operational cost, and penalty for deviation converge gradually and reach a limit value as the iteration number increases. Therefore, the proposed SAM-SAC algorithm effectively handles the VPP regulation service disaggregation problem.

C. Comparisons Between the Proposed SAM-SAC Algorithm and the Existing DRL Algorithms

The computational efficiency comparison between the proposed SAM-SAC algorithm and the existing algorithms are conducted. 10 instances are trained using different DRL algorithms. The average single-step reward of every 100 environment steps in the training process is shown in Fig. 5. In Fig. 5 and Fig. 6, the solid line corresponds to the average step-wise reward of every 100 steps over 10 trials and the shaded region is bounded by the minimum and maximum step-wise rewards of every 100 steps over 10 trials.

SAC is a promising DRL algorithm in state-of-the-art literature [18], [20]. SAC essentially learns a stochastic policy to maximize the expected reward and entropy simultaneously, which helps the agent to encourage exploration to avoid premature convergence. The proposed SAM-SAC algorithm inherits such an advantage and achieves comparable computational efficiency compared with the traditional SAC proposed in [20], as shown in Fig. 5. Therefore, the proposed SAM-SAC algorithm outperforms the most efficient deterministic policy gradient algorithm named deep deterministic policy gradient (DDPG) [41].

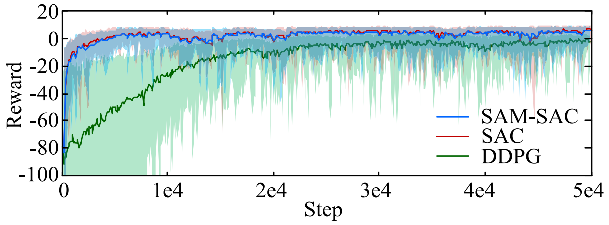


Fig. 5. Convergence comparison of SAM-SAC, SAC, and DDPG.

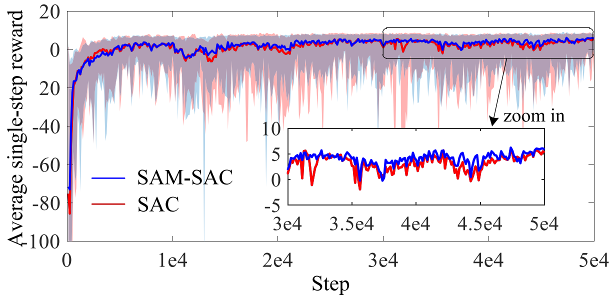


Fig. 6. Convergence comparison of the SAM-SAC algorithm and the SAC algorithm under the reward noise in the range of [−2.0, 2.0].

Furthermore, compared with the traditional SAC algorithm, the proposed SAM-SAC algorithm has significant superiorities in terms of robustness and applicability, i.e., providing higher robustness to the reward noise and more adaptive to the environmental parameter changing. Numerical simulations to verify these advantages are presented as follows.

1) Robustness of the proposed SAM-SAC algorithm: Due to the telemetry error and parameter inaccuracy of the practical environment, the calibration of the environment reward is inaccurate in reality. To imitate the difference between the reward in the algorithm computation and the reward in the real world, the random noises with uniform distribution in the range of [−2.0, 2.0] are added on the basis of the environmental reward. 10 trials with different random seeds and training datasets are trained using the proposed SAM-SAC algorithm and the SAC algorithm under the reward noise condition. The average single-step rewards of every 100 environment steps are presented in Fig. 6.

According to [21] and [22], SAM method can retain the robustness to label noise and enhance the training accuracy across various noise level. The proposed SAM-SAC algorithm inherits such an advantage of SAM method. The simulation results reveal that the SAM-SAC algorithm can obtain a higher reward than traditional SAC under the environmental noise condition. Moreover, to further illustrate the superiority of the proposed algorithm, the average computational results of the last 1e4 steps in the iteration are presented in Table IV. Compared with the traditional SAC, the proposed SAM-SAC algorithm can enhance the regulation request tracking error, reduce the VPP operational cost, and relieve the penalty for the deviation between the regulation trajectories and the intraday dispatch plans. Therefore, the SAM-SAC algorithm

TABLE IV
COMPUTATIONAL RESULT COMPARISONS OF THE LAST 1E4 STEPS USING DIFFERENT ALGORITHMS

Range of the reward noise	[-1.0, 1.0]		[-2.0, 2.0]	
Algorithm	SAM-SAC	SAC	SAM-SAC	SAC
Average single-step reward	4.4880	3.7711	4.3203	3.4705
Average single-step penalty for deviation	3.1571	3.6264	3.1807	3.679
Average absolute value of single-step tracking error (MW)	0.0924	0.1071	0.1032	0.1230
Average single-step operational cost ($)	0.2862	0.3063	0.2934	0.3241

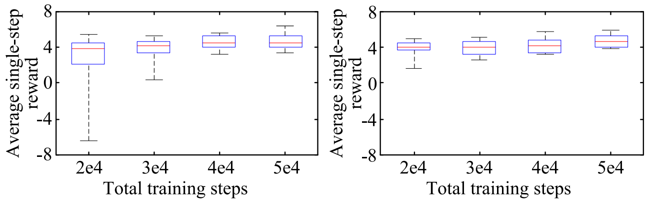
contributes to notable improvement in providing robustness to the reward noise.

2) Applicability of the proposed SAM-SAC algorithm: The proposed approach is still available if the offline simulator has few data about the response of DERs, since the DRL method does not rely on the exact environmental model and parameters. In that case, the dynamic parameters of these DERs' can be sampled according to the probability distribution from historical statistics. Due to the huge number of small-capacity DERs, the absence of partial DER parameters will not affect the overall implementation performance of the proposed approach. To reflect the parameter deviation between the practical environment and the offline simulation environment, 20% random noise with uniform distribution is added to the approximation parameters in the offline simulator (including the dynamic parameters and operation cost coefficients of FR aggregators) before each trial. In addition, in the imitated practical environment, the FR devices are assigned with a certain probability of not following the VPP regulation command, which is represented by the uncontrollable probability presented in Table I. The training process is conducted in the offline simulator first, and then the obtained control policies are tested in the imitated practical environment.

Twenty trials are conducted with different random seeds and training datasets. The comparison between the SAC algorithm and the proposed SAM-SAC algorithm in terms of the average single-step reward in the practical environment is presented in Fig. 7. According to the numerical simulations, it can be indicated that, even there exists parameter deviation between the practical and simulation environment, the performance of the proposed approach is still decent. Moreover, it can be observed that the algorithm performance in the practical environment can be improved as the total training step grows.

Furthermore, the deviation of the reward obtained by the proposed SAM-SAC algorithm is smaller than the traditional SAC. Therefore, compared with the SAC algorithm, the proposed SAM-SAC algorithm is more adaptive to environmental parameter changing. This merit indicates that the SAM-SAC algorithm has promising application potential in the proposed two-stage VPP disaggregation problem since inevitable discrepancies always exist between the simulation environment and the practical environment.

To facilitate the DRL approach to learn the characteristics of the environment, the characteristics of the sub-aggregators should be incorporated into the training environment, as presented in Section III. The characteristics can be divided into the following two categories. The first category is the regulation range constraints in (11)-(14) under power and energy capacity limitations. It is necessary to learn the constraints determining the regulation range, because regulation range affects the exploration space of the DRL algorithm and determines whether the obtained solution is feasible to be conducted. The second category is the dynamic model depicting the dynamic characteristics of the aggregators, as in (10). The following experiments verify the significance and necessity of learning the dynamic model of the sub-aggregators.



(a). SAC algorithm.
(b). SAM-SAC algorithm.

Fig. 7. Comparison between the SAC and the proposed SAM-SAC in terms of the average single-step reward in the practical environment.

TABLE V
COMPARISON OF THE AVERAGE SINGLE-STEP REWARD IN THE PRACTICAL SYSTEM UNDER THE TWO APPROACHES

Item		Case#I	Case#II
Parameters in the practical environment	Inertia	H_i	$10H_i$
	Time delay	T_i	$5T_i$
Average single-step reward	Proposed approach	4.887	3.675
	Compared approach	3.111	1.880
Average absolute value of the single-step tracking error (MW)	Proposed approach	0.092	0.233
	Compared approach	0.139	0.441

To verify the significance of learning the dynamic model, a simpler method is adopted for comparison, in which the DRL approach learns to split the reference regulation command without considering the dynamic model of sub-aggregators in the offline training stage. The compared approach and the proposed approach are employed to train the disaggregation policy in the offline simulator first and then tested in the practical environment, respectively. Without loss of generality, two scenarios with different dynamic parameters (inertia and time delay) are configured to compare the performance of different controllers. 10 different trials with different random seeds and training datasets are conducted separately. The experimental results are presented in Table V. The average single-step reward and the average absolute value of the tracking error are computed based on the test dataset in the imitated practical environment. Moreover, to visualize the advantage of the proposed approach, the regulation command tracking accuracy comparisons using the two approaches in Case#II is displayed in Fig. 8. The experimental results show that the proposed approach can improve the reward and reduce the tracking error effectively compared with the simpler method. Therefore, it is necessary to learn the dynamic model of the aggregators using the proposed DRL approach.

D. Advantage of the Proposed Two-Stage DRL Approach

After experience accumulation in the offline-stage, the trained policy network and Q network can be transferred to

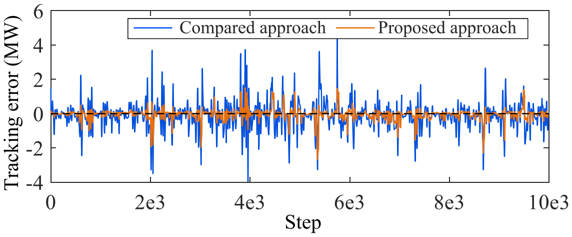


Fig. 8. Comparison of the VPP tracking accuracy using different approaches.

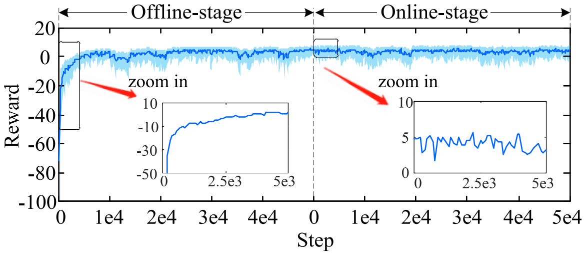


Fig. 9. Reward of the offline-stage and the online-stage.

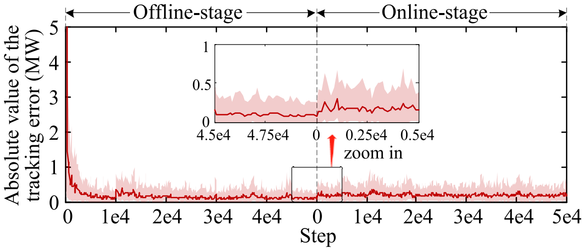


Fig. 10. Absolute value of the tracking error of the offline-stage and the online-stage.

TABLE VI
PERFORMANCE COMPARISONS OF THE ONLINE START-UP PROCESS USING DIFFERENT APPROACHES

Item	Single-stage DRL approach	Two-stage DRL approach
Average single-step reward	-6.385	4.152
Average single-step penalty for Deviation	9.187	2.515
Average single-step tracking error (MW)	0.513	0.195
Accumulated operational cost ($)	2070	1385

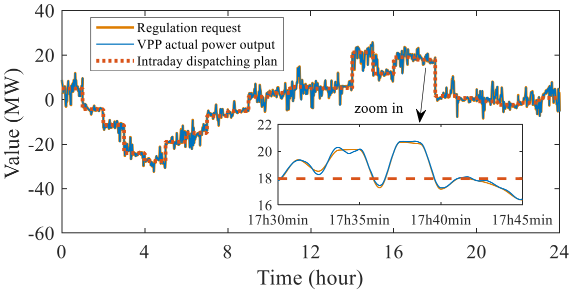


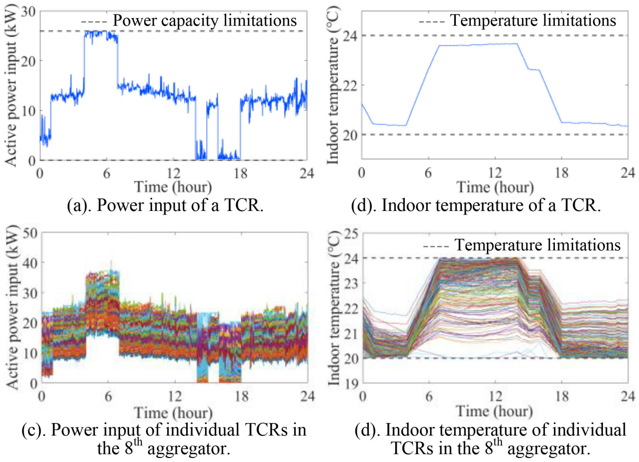
Fig. 11. Regulation request and actual power output of the VPP.

online implementation. Before conducting each trial, to imitate the randomness of individual DERs in the practical environment, each DER is assigned with an uncontrollable probability presented in Table I in the online environment, and 20% random noise with uniform distribution is added to the approximation parameters in the offline simulator. The reward and tracking error of the offline-stage and the online-stage are shown in Fig. 9 and Fig. 10, in which the solid line corresponds to the average step-wise returns (reward and tracking error) of every 100 steps over 10 trials and the shaded region is bounded by the minimum and maximum step-wise returns of every 100 steps over 10 trials.

It can be indicated from Fig. 9 that the deviation of the system reward during the online start-up process is significantly reduced compared with that of the offline start-up process. In addition, the tracking error of the online start-up process still can be maintained within an acceptable range, as shown in Fig. 10. Therefore, the prior knowledge learned from the developed offline simulator could help promote the algorithms to take satisfactory actions at the initial online learning process and improve the performance of the start-up process of online implementation.

The comparisons of the online start-up process (initial 5000 simulation steps) using the single-stage DRL approach and the proposed two-stage DRL approach are presented in Table VI. The reason for only comparing the performance of the initial 5000 steps is that: the results of the initial 5000 steps can evidently reveal the advantages of the two-stage DRL approach in improving the algorithm's performance during the online start-up process. The computational results further illustrate the superiority of the proposed two-stage DRL approach in terms of higher tracking accuracy, lower operational cost, and smaller deviation between the regulation trajectories and the intraday dispatch plans. Therefore, the expensive operational cost and tracking error of the practical power systems in the online start-up process can be alleviated using the proposed two-stage DRL approach.

The requirements of the power system in real-time operation can be met in terms of computational efficiency, regulation performance, and economy. (1) During real-time operation, the policy network directly generates the regulation commands. Statistics show that the average time of computing the action variables by the neural networks is nearly 0.79 milliseconds, which is much less than the control interval (typically several seconds) of regulation command issued by ancillary services market operator [2], [3]. (2) After 5e4 consecutive iterations of the online training stage, the percentage of the tracking error (ratio of the tracking error to the reference command) is decreased to 0.87%, which meets the tracking error threshold (1.5%) of Guangdong provincial power market in China Southern Power Grid. (3) According to Table VI, the proposed two-stage approach has the advantage of reducing the VPP operational cost compared with the single-stage approach. With the continuous updating of the control policy, the economy of the disaggregation method is further improved, which can effectively promote the long-term economical operation of the VPP. To visualize the application result of the proposed approach in real-time operation, the aggregate power output trajectory in a typical day of the VPP is presented in Fig. 11. It can be indicated that the aggregate power trajectory of the VPP can track the regulation request issued by the superior power



(a). Power input of a TCR.

(d). Indoor temperature of a TCR.

(c). Power input of individual TCRs in the 8th aggregator.

(d). Indoor temperature of individual TCRs in the 8th aggregator.

Fig. 12. Disaggregation results of individual DERs.

system operator accurately by regulating the electric power of different controllable devices cooperatively. Therefore, the proposed method is effective to promote the VPP to provide frequency regulation service for power systems.

Due to page limitations, the disaggregation results of some representative DERs and aggregators (the 8th aggregator) are presented in Fig. 12. The daily power input curve and indoor temperature curve of a thermal controllable resident (TCR) are presented Fig. 12(a) and (b), and the daily power input curves and indoor temperature curves of all the TCRs in the 8th aggregator (composed of 273 TCRs) are shown in Fig. 12(c) and (d), respectively. The simulation results reveal that the indoor temperatures of the individual TCRs can be controlled within the predetermined comfort range ($20°C \sim 24°C$) by adjusting the power consumption of the TCRs within the capacity limits. Therefore, the disaggregation solutions obtained by the proposed DRL approach can be maintained within the feasible region of the DERs.

VI. Conclusion

In this study, a two-stage DRL based frequency regulation service disaggregation approach for a VPP is proposed, through which the time-varying regulation request issued by the superior power system can be decomposed to DER aggregators efficiently in the inaccurate environmental model. The developed two-stage DRL approach can improve the performance of the online start-up process with prior knowledge. The proposed SAM-SAC algorithm is adaptive to environmental parameters changes and exhibits robustness to the reward noise. Numerical simulation results illustrate that the proposed approach can manage various controllable resources economically and track the regulation requests with high accuracy, which can promote the VPP to provide regulation service for power systems.

Appendix
Proof of the Two-Step Iteration to Solve the Problem in (35)

To solve the adaptive sharpness-aware minimization problem presented in (35). The optimal value of ε is computed in the inner layer and the optimal value of θ is computed in the outer layer. By introducing the auxiliary variable $\tilde{\varepsilon} = T_\theta^{-1}\varepsilon$, the min-max problem in (35) is reformulated as follows:

$$\min_{\theta}\max_{\|\tilde{\varepsilon}\|_2\le\rho} J_Q(\theta+T_\theta\tilde{\varepsilon})+\frac{\lambda}{2}\|\theta\|_2^2 \tag{A1}$$

Via the first-order Taylor expansion of $J_Q(\theta)\big|_{\theta=\theta_k}$, the approximation of the optimal value of the inner layer variable $\tilde{\varepsilon}$ is derived as follows:

$$\begin{aligned}\tilde{\varepsilon}_k^* &= \arg\max_{\|\tilde{\varepsilon}\|_2\le\rho} J_Q\big(\theta_k+T_{\theta_k}\tilde{\varepsilon}\big)\\ &\approx \arg\max_{\|\tilde{\varepsilon}\|_2\le\rho}\Big(J_Q(\theta_k)+\tilde{\varepsilon}^\top T_{\theta_k}\nabla J_Q(\theta_k)\Big)\\ &= \arg\max_{\|\tilde{\varepsilon}\|_2\le\rho}\tilde{\varepsilon}^\top T_{\theta_k}\nabla J_Q(\theta_k).\end{aligned}\tag{A2}$$

According to the conclusion drawn from [21], the solution of (A2) can be formulated as follows:

$$\begin{aligned}\tilde{\varepsilon}_k^* &= \rho\,\mathrm{sign}\Big(T_{\theta_k}\hat{\nabla}J_Q(\theta_k)\Big)\frac{\Big|T_{\theta_k}\hat{\nabla}J_Q(\theta_k)\Big|}{\Big\|T_{\theta_k}\hat{\nabla}J_Q(\theta_k)\Big\|_2}\\ &= \rho\frac{T_{\theta_k}\hat{\nabla}J_Q(\theta_k)}{\Big\|T_{\theta_k}\hat{\nabla}J_Q(\theta_k)\Big\|_2}\end{aligned}\tag{A3}$$

where $|\cdot|$ denotes element-wise absolute value; $\|\cdot\|_2$ denotes the 2-norm; $\mathrm{sign}(\cdot)$ denotes the signum function of a vector; $T_{\theta_k} = \mathrm{diag}[|\theta_k(1)|,\ldots,|\theta_k(m)|]$ is the element-wise normalization operator satisfying scale-invariant property $T_{A\theta_k}^{-1}A = T_{\theta_k}^{-1}$; A is the invertible scaling operator; $\theta_k = [\theta_k(1),\ldots,\theta_k(m)]$; m is the number of neural network parameters.

Substituting $\varepsilon_k^* = T_{\theta_k}\tilde{\varepsilon}_k^*$ into (A3), ε_k^* can be obtained:

$$\varepsilon_k^* = T_{\theta_k}\tilde{\varepsilon}_k^* = \rho\frac{T_{\theta_k}^2\hat{\nabla}J_Q(\theta_k)}{\Big\|T_{\theta_k}\hat{\nabla}J_Q(\theta_k)\Big\|_2}.\tag{A4}$$

Via the first-order Taylor expansion of $J_Q(\theta)|_{\theta=\theta_k+\varepsilon_k}$, the approximation of the optimal value of the outer layer variable θ can be derived. Afterwards, by introducing a step size δ_Q in the process, the optimal value of θ can be solved approximately by using the gradient descent method, as derived in (A5).

$$\begin{aligned}\theta_{k+1}^* &= \arg\min_{\theta} J_Q(\theta+\varepsilon_k)+\frac{\lambda}{2}\|\theta\|_2^2\\ &\approx \arg\min_{\theta}\Big(J_Q(\theta_k+\varepsilon_k)+(\theta-\theta_k)^\top\hat{\nabla}J_Q(\theta_k+\varepsilon_k)+\frac{\lambda}{2}\|\theta\|_2^2\Big)\\ &= \arg\min_{\theta}\Big((\theta-\theta_k)^\top\hat{\nabla}J_Q(\theta_k+\varepsilon_k)+\frac{\lambda}{2}\|\theta\|_2^2\Big)\\ &\approx \theta_k-\delta_Q\Big(\hat{\nabla}J_Q(\theta_k+\varepsilon_k)+\lambda\theta_k\Big).\end{aligned}\tag{A5}$$

In summary, the two-step iteration to solve the min-max adaptive sharpness-aware minimization problem is as follows:

$$\begin{cases}\varepsilon_k = \rho\frac{T_{\theta_k}^2\nabla J_Q(\theta_k)}{\|T_{\theta_k}\nabla J_Q(\theta_k)\|_2}\\ \theta_{k+1} = \theta_k-\delta_Q\Big(\hat{\nabla}J_Q(\theta_k+\varepsilon_k)+\lambda\theta_k\Big).\end{cases}\tag{A6}$$

References

[1] M. P. Evans, S. H. Tindemans, and D. Angeli, "A graphical measure of aggregate flexibility for energy-constrained distributed resources," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 106–117, Jan. 2020.

[2] "PJM Manual 11: Energy & Ancillary Services Market Operations." [Online]. Available: https://www.pjm.com/-/media/documents/manuals/m11-redline.ashx (Accessed: Jan. 5, 2021).

[3] California ISO: Pay for Performance Regulation Draft Final Proposal." Feb. 2012. [Online]. Available: http://www.caiso.com/Documents/DraftFinalProposal-PayforPerformanceRegulation.pdf#search=Performance%20Regulation (Accessed: Feb. 13, 2012).

[4] T. Morstyn, N. Farrell, S. J. Darby, and M. D. McCulloch, "Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants," *Nat. Energy*, vol. 3, no. 2, pp. 94–101, Feb. 2018.

[5] S. Babaei, C. Zhao, and L. Fan, "A data-driven model of virtual power plants in day-ahead unit commitment," *IEEE Trans. Power Syst.*, vol. 34, no. 6, pp. 5125–5135, Nov. 2019.

[6] J. Naughton, H. Wang, M. Cantoni, and P. Mancarella, "Co-optimizing virtual power plant services under uncertainty: A robust scheduling and receding horizon dispatch approach," *IEEE Trans. Power Syst.*, vol. 36, no. 5, pp. 3960–3972, Sep. 2021.

[7] A. Baringo, L. Baringo, and J. M. Arroyo, "Day-ahead self-scheduling of a virtual power plant in energy and reserve electricity markets under uncertainty," *IEEE Trans. Power Syst.*, vol. 34, no. 3, pp. 1881–1894, May 2019.

[8] T. Zhang, S. X. Chen, H. B. Gooi, and J. M. Maciejowski, "A hierarchical EMS for aggregated BESSs in energy and performance-based regulation markets," *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 1751–1760, May 2017.

[9] Q. Shi, F. Li, Q. Hu, and Z. Wang, "Dynamic demand control for system frequency regulation: Concept review, algorithm comparison, and future vision," *Elect. Power Syst. Res.*, vol. 154, pp. 75–87, Jan. 2018.

[10] E. Dall'Anese, S. S. Guggilam, A. Simonetto, Y. C. Chen, and S. V. Dhople, "Optimal regulation of virtual power plants," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 1868–1881, Mar. 2018.

[11] L. A. D. Espinosa, A. Khurram, and M. Almassalkhi, "Reference-tracking control policies for packetized coordination of heterogeneous DER populations," *IEEE Trans. Control Syst. Technol.*, vol. 29, no. 6, pp. 2427–2443, Nov. 2021, doi: 10.1109/TCST.2020.3039492.

[12] J. L. Mathieu, S. Koch, and D. S. Callaway, "State estimation and control of electric loads to manage real-time energy imbalance," *IEEE Trans. Power Syst.*, vol. 28, no. 1, pp. 430–440, Feb. 2013.

[13] G. He, Q. Chen, C. Kang, P. Pinson, and Q. Xia, "Optimal bidding strategy of battery storage in power markets considering performance-based regulation and battery cycle life," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2359–2367, Sep. 2016.

[14] M. Abouheaf, Q. Gueaieb, and A. Sharaf, "Load frequency regulation for multi-area power system using integral reinforcement learning," *IET Gener. Transm. Distrib.*, vol. 13, no. 19, pp. 4311–4323, 2019.

[15] Y. Ye, D. Qiu, M. Sun, D. Papadaskalopoulos, and G. Strbac, "Deep reinforcement learning for strategic bidding in electricity markets," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1343–1355, Mar. 2020.

[16] E. Foruzan, L.-K. Soh, and S. Asgarpoor, "Reinforcement learning approach for optimal distributed energy management in a microgrid," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 5749–5758, Sep. 2018.

[17] Y. Ye, Y. Tang, H. Wang, X.-P. Zhang, and G. Strbac, "A scalable privacy-preserving multi-agent deep reinforcement learning approach for large-scale peer-to-peer transactive energy trading," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 5185–5200, Nov. 2021, doi: 10.1109/TSG.2021.3103917.

[18] H. Liu and W. Wu, "Two-stage deep reinforcement learning for inverter-based volt-VAR control in active distribution networks," *IEEE Trans. Smart Grid*, vol. 12, no. 3, pp. 2037–2047, May 2021.

[19] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

[20] T. Haarnoja *et al.*, "Soft actor-critic algorithms and applications," Jan. 2019, *arXiv:1812.05905*.

[21] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *Proc. Int. Conf. Learn. Represent.*, Vienna, Austria, May 2021, pp. 1–20.

[22] J. Kwon, J. Kim, H. Park, and I. K. Choi, "ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, Jul. 2021, pp. 5905–5914.

[23] Z. Yi, Y. Xu, W. Gu, L. Yang, and H. Sun, "Aggregate operation model for numerous small-capacity distributed energy resources considering uncertainty," *IEEE Trans. Smart Grid*, vol. 12, no. 5, pp. 4208–4224, Sep. 2021.

[24] M. Mohammadi, M. M. Arefi, P. Setoodeh, and O. Kaynak, "Optimal tracking control based on reinforcement learning value iteration algorithm for time-delayed nonlinear systems with external disturbances and input constraints," *Inf. Sci.*, vol. 554, pp. 84–98, Apr. 2021.

[25] K. S. Ko, S. Han, and D. K. Sung, "A new mileage payment for EV aggregators with varying delays in frequency regulation service," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 2616–2624, Jul. 2018.

[26] Z. Wang, W. Wu, and B. Zhang, "A distributed quasi-newton method for droop-free primary frequency control in autonomous microgrids," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 2214–2223, May 2018.

[27] A. Sadeghi-Mobarakeh and H. Mohsenian-Rad, "Optimal bidding in performance-based regulation markets: An MPEC analysis with system dynamics," *IEEE Trans. Power Syst.*, vol. 32, no. 2, pp. 1282–1292, Mar. 2017.

[28] L. Subramanian, V. Debusschere, H. B. Gooi, and N. Hadjsaid, "A distributed model predictive control framework for grid-friendly distributed energy resources," *IEEE Trans. Sustain. Energy*, vol. 12, no. 1, pp. 727–738, Jan. 2021.

[29] X. Cao, B. Stephen, I. F. Abdulhadi, C. D. Booth, and G. M. Burt, "Switching Markov Gaussian models for dynamic power system inertia estimation," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3394–3403, Sep. 2016.

[30] J. Zhang and H. Xu, "Online identification of power system equivalent inertia constant," *IEEE Trans. Ind. Electron.*, vol. 64, no. 10, pp. 8098–8107, Oct. 2017.

[31] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[32] S. M. Ross, *Stochastic Processes*. New York, NY, USA: Wiley, 1996, pp. 231–249.

[33] X. Chen, G. Qu, Y. Tang, S. Low, and N. Li, "Reinforcement learning for decision-making and control in power systems: Tutorial, review, and vision," Jul. 2021, *arXiv:2102.01168*.

[34] K. S. Ko, S. Han, and D. K. Sung, "Performance-based settlement of frequency regulation for electric vehicle aggregators," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 866–875, Mar. 2018.

[35] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.

[36] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *Proc. Int. Conf. Learn. Represent.*, San Juan, Puerto Rico, May 2016, pp. 1–21.

[37] C. Wang and K. Ross, "Boosting soft actor-critic: Emphasizing recent experience without forgetting the past," 2019, *arXiv:1906.04009*.

[38] A. Gorbunov, A. Dymarsky, and J. Bialek, "Estimation of parameters of a dynamic generator model from modal PMU measurements," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 53–62, Jan. 2020.

[39] R. Garrido and A. Concha, "Inertia and friction estimation of a velocity-controlled servo using position measurements," *IEEE Trans. Ind. Electron.*, vol. 61, no. 9, pp. 4759–4770, Sep. 2014.

[40] Z. Yi, Dec. 2021, "Supplemental Materials of DER Parameters," Figshare. [Online]. Available: https://doi.org/10.6084/m9.figshare.17705633.v1

[41] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learn. Represent.*, San Juan, Puerto Rico, May 2016, pp. 1–14.

Zhongkai Yi (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the Harbin Institute of Technology in 2016 and 2018, respectively, and the Ph.D. degree in electrical engineering from Tsinghua University in January 2022.

He is currently an Algorithm Expert with the Decision Intelligence Lab, Machine Intelligence Research Sector, Alibaba DAMO Academy. His research interests include the optimization and machine learning in power systems. He was a recipient of the Advanced Technology Talent Program "AliStar" of Alibaba.

Yinliang Xu (Senior Member, IEEE) received the B.S. and M.S. degrees in control science and engineering from the Harbin Institute of Technology, China, in 2007 and 2009, respectively, and the Ph.D. degree in electrical and computer engineering from New Mexico State University, Las Cruces, NM, USA, in 2013.

He is currently an Associate Professor with Tsinghua–Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School, Tsinghua University. His research interests include distributed control and optimization of power systems, renewable energy integration, and microgrid modeling and control. He is an Associate Editor for the *IET Renewable Power Generation*, *IET Smart Grid*, and *IET Generation, Transmission & Distribution*.

Xue Wang received the B.S. degree in industrial engineering from Tsinghua University in 2013, and the Ph.D. degree in industrial engineering and operations research from Pennsylvania State University in 2019.

He is currently a Senior Algorithm Expert with the Decision Intelligence Lab, Alibaba DAMO Academy. His research interests include statistical learning, optimization, and stochastic modeling.

Wei Gu (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Southeast University, China, in 2001 and 2006, respectively.

He is currently a Professor with the School of Electrical Engineering, Southeast University. He is the Director of the Institute of Distributed Generations and Active Distribution Networks. His research interests include distributed generations and microgrids, and integrated energy systems.

Hongbin Sun (Fellow, IEEE) received the double B.S. degrees from Tsinghua University in 1992, and the Ph.D. degree from the Department of Electrical Engineering, Tsinghua University in 1996.

He is currently a ChangJiang Scholar Chair Professor and the Director of Energy Management and Control Research Center, Tsinghua University. His technical areas include electric power system operation and control with specific interests on the energy management system, automatic voltage control, and energy system integration. He also serves as the Editor for the IEEE TRANSACTIONS ON SMART GRID, an Associate Editor of *IET Renewable Power Generation*, and a member of the Editorial Board of four international journals and several Chinese journals.

Qiuwei Wu (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees in power system and its automation from the Nanjing University of Science and Technology in 2000 and 2003, respectively, and the Ph.D. degree in power system engineering from Nanyang Technological University, Singapore, in 2009.

He has been a Tenured Associate Professor with Tsinghua–Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School, Tsinghua University, since January 2022. His research interests are decentralized/distributed optimal operation and control of power systems and integrated energy systems.

Chenyu Wu (Member, IEEE) received the Ph.D. degree in electrical engineering from Southeast University, Nanjing, China, in 2019.

He is currently a Postdoctoral Research Fellow with the School of Electrical Engineering, Southeast University. His research interests include integrated energy system, game theory, energy market, and decentralized optimization.