IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON KEY ENABLING TECHNOLOGIES FOR PROSUMER ENERGY MANAGEMENT

Received September 30, 2020, accepted October 4, 2020, date of publication October 6, 2020, date of current version October 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029189

Virtual Generation Alliance Automatic Generation Control Based on Deep Reinforcement Learning

JIAWEN LI AND TAO YU

College of Electric Power, South China University of Technology, Guangzhou 510640, China

Corresponding author: Tao Yu (taoyul@scut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 51777078.

ABSTRACT This article proposes a distributed hierarchical automatic generation control (AGC) framework with multiple regulation units in the performance-based frequency regulation market, named virtual generation alliance automatic generation control (VGA-AGC), aiming to achieve the coordination of control algorithm and AGC dispatch algorithm and adapt to the development trend of AGC from centralized framework to centralized-decentralized framework. The framework also involves a multi agent distributed multiple improved deep deterministic policy gradient (MADMI-TD3) algorithm that is characterized by excellent global search capability and optimizing speed. The algorithm can help create an optimal AGC strategy in a randomization environment so as to obtain an optimal cooperative control of AGC. According to a simulation verification on the LFC model for an interconnected power grid of a province, the algorithm is superior to the current algorithms and conventional engineering methods in terms of control performance and economic benefits. In other words, the algorithm can improve control performance and reduce the regulation mileage payment.

INDEX TERMS Performance-based frequency regulation market, virtual generation alliance automatic generation control (VGA-AGC), multi agent distributed multiple improved deep deterministic policy gradient, regulation mileage payment, centralized-decentralized autonomy.

I. INTRODUCTION

The ever-increasing innovation of renewable energy makes the power grid more dispersed, diverse, and random [1]–[3]. A traditional AGC strategy has difficulty dealing with the strong random disturbance. In the strategy, the total AGC generation power command of system is generated and dispatched to units through the proportion integration (PI) controller and the proportional dispatch method [4].

Especially when there is a sudden power disturbance in a complicated power grid system, in which the large-scale wind turbine has a poor disturbance tolerance, the traditional AGC strategy may result in chain reactions, which can cause a larger power disturbance and affect the safety as well as stability of the system frequency. For example, during the ''8·9'' blackout in the UK [5] and ''9.28'' blackout in Australia [6], the major power failure accidents were all caused by the off grid of wind turbine, which resulted in serious reduction of frequency. The AGC could not timely be responded and the frequency regulation capacity was short during the accident. Thus, it is very important to improve the response speed and control performance of AGC in a complicated power grid system with large-scale wind turbine.

The algorithms can generally be divided into two categories. The first category is the control algorithm such as the conventional PID algorithm [7], [8], sliding mode control (SMC) [9], active disturbance rejection control (ADRC) [10], fractional Order PID (FOPID) [11], fuzzy control [8], and reinforcement learning series such as Q learning algorithm [12], [13], Q learning algorithm [14], R(λ) learning algorithm [15], (Deep Q-Network) DQN [16], and (Double Deep Q-Network)DDQN [17]. Generally speaking, these algorithms take the entire power grid as a single area for calculation of generation command, which is then proportionally distributed to AGC regulation units. The other category of algorithms refers to generation power command dispatch algorithm such as classical genetic algorithm (GA), quadratic programming, gray wolf optimizer (GWO) [18]–[20], proportional algorithm, particle swarm optimization (PSO) [21], moth-flame optimization (MFO) [22], whale optimization algorithm (WOA) [23], ant lion optimizer (ALO) [24], dragonfly algorithm (DA) [25], group search optimizer (GSO) [26], chicken swarm optimization (CSO) [27], sine cosine algorithm, SCA) [28], and etc [28], [29]. The classical

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

PID control algorithm is generally adopted as the control algorithm and the dispatch algorithm is generally used to dispatch the total power regulation command of ACG to each AGC unit, aiming to minimize the regulation payment. The separation of the two types of algorithms has certain advantages. For example, the control algorithm and dispatch algorithm can separately be designed. The two types of algorithms also have a problem in terms of cooperation. The control algorithm aims to minimize the control deviation of frequency while dispatch algorithm the regulation payment. The combination of two types can reduce the frequency deviation and the regulation payment, thus improving the control performance as well as lowering the regulation payment of the AGC. The methods, mentioned above, are all based on the centralized control framework. It is necessary to collect real time operation data from all units, which means a large amount of information for transmission. When the sizes of units increase, the convergence time of the above methods can greatly be improved, however, it may become difficult to meet the real-time control requirements of AGC [28].

The performance-based frequency regulation market (hereinafter referred to as ''frequency regulation market'') [30] is proposed in Order No. 755, issued by the Federal Energy Regulatory Commission (FERC) in 2011, aiming to encourage more fast- response regulation units, such as wind turbine unit, photovoltaic generation unit, and flexible loads, to participate in the secondary frequency regulation. Due to the new frequency regulation market mechanism, AGC regulation payment is changed from the original simple fixed payment per unit of regulation output to the dynamic compensation payment influenced by the comprehensive frequency regulation performance index, frequency regulation mileage, and frequency regulation mileage quotation. The original combination of control algorithm and dispatch algorithm (hereinafter referred to as the conventional combined algorithm) of AGC scheduling framework cannot be suitable to the new frequency regulation market mechanism, and the problem concerning coordination of control algorithm and dispatch algorithm has become more serious.

This article aims to design a virtual generation alliance automatic generation control (VGA-AGC) framework with various units including distributed energy units and flexible load units in order to solve the problem concerning coordination of control algorithm and dispatch algorithm in performance-based frequency regulation, make AGC to control more frequency regulation units, and adapt to the development trend of AGC from centralized framework to centralized-decentralized framework. The VGA-AGC framework also involves the coordination of AGC control algorithm and generation power command dispatch algorithm. The VGA-AGC framework, which is based on MADMI-TD3 algorithm, has following characteristics.

1) The proposed MADMI-TD3 algorithm employs different parameters of multiple actor networks and critic networks for distributed optimizing. A few techniques such as classified experience replay, variable noise models, and warm boot of experience pool are used to improve the global search ability and optimizing speed of the algorithm. The algorithm can obtain the optimal control strategy under strong random environment with random disturbance caused by large-scale distributed energy in the power grid.

2) The proposed MADMI-TD3 algorithm employs different parameters of multiple actor networks and critic networks for distributed optimizing, in addition, several techniques like classified experience replay, variable noise models, warm boot of experience pool are utilized to obtain an adaptive reinforcement learning control algorithm with superior global search ability and optimizing speed. The algorithm can obtain the optimal control strategy under strong random environment with random disturbance caused by large-scale distributed energy introduction in the power grid.

3) A simulation verification on the LFC model for an interconnected power grid of a province has shown that the algorithm proposed here is superior to the current algorithms and conventional engineering methods in terms of control performance and economic benefits: that is, the improvement of the control performance and reduction of the regulation mileage payment.

II. VIRTUAL GENERATION ALLIANCE AUTOMATIC GENERATION CONTROL(VGA-AGC)

A. CONTROL FRAMEWORK

Different from conventional AGC system, VGA–AGC has a framework with multiple agents for generating and dispatching of the AGC total generation power command. Agents, composing the virtual generation alliance, cooperate with each other in all the layers. The AGC control cycle is 4s.

B. VIRTUAL GENERATION ALLIANCE

Virtual generation alliance: Professor Clerc divided the whole particle swarm into several subgroups, namely ''alliances'' [31], and each ''alliance'' consists of several particles. Hence, this article divided the units into several territory groups of units according to their type, namely ''alliance''. It is actually a new dispatch and control layer added between the center scheduling and plant controller (PLC) as a form of centralized-decentralized autonomy, corresponding to territory groups of units. For each territory, there is an administrator: Lord agent. As shown in Figure 1, for the VGA-AGC framework, four roles corresponding to the agents are proposed, including king agent, general agent, lord agent, and knight (units).

1) KING AGENT

It refers to the controller for an area of the power grid. In this article, the king agent based on MADMI-TD3 replaced the conventional PI controller. In the process of offline training, the agent can observe the state of each agent in the environment, so as to evaluate the action made by the king agent, thus adjusting its own actions based on global information. As compared to conventional controller which

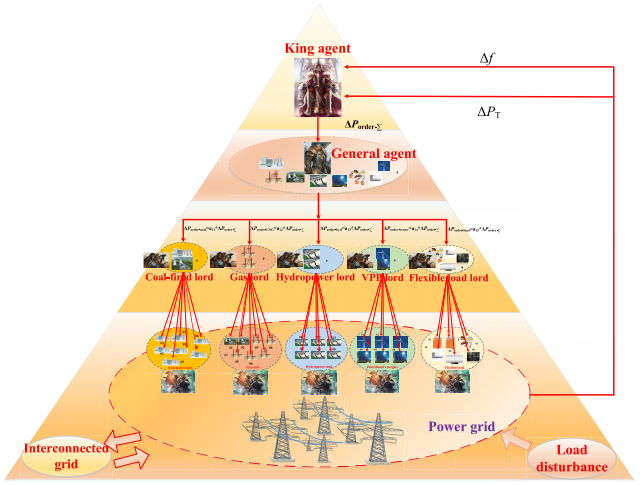

FIGURE 1. VGA-AGC framework based on MADMI-TD3.

makes decisions only based on area control error (ACE), the king agent has superior robustness and coordination. The king agent is responsible for the real-time output of total power regulation command.

2) GENERAL AGENT

It is the total dispatch agent inferior to the king agent. It dispatches the total power regulation command issued by the king agent to the next level of agent: lord agent. General agent also needs to observe the state of all agents in offline training process, so as to evaluate and adjust its own actions, thus obtaining the optimal dispatch strategy.

3) LORD AGENT

Different types of units are classified in many groups which named as the territory groups. The agent responsible for the territory groups of units is lord agent, and it is responsible for dispatching the generation power command issued by the general agent to units which are in their territory groups. Lord agents includes the following types: Coal lord for coal-fired generation unit, gas lord for CHP and Liquefied Natural Gas (LNG) units, hydropower lord for hydroelectric unit, flexible load,virtual power plant (VVP) lord for various distributed energies, such as wind power, photovoltaic power, P2G and etc.

4) KNIGHT (UNITS)

It refers to the generation units which is responsible for outputting power under the generation command of the lords.

C. APPLICATION PROCESS

Before applying the VGA-AGC to the power grid, which is called online testing, it needs to participate sufficient offline training:

1) OFFLINE TRAINING

During offline training, in each of the AGC control cycle, each agent communicates with the EMS system of scheduling center, and at the same time, all actions performed by other agents can be observed by the agent, so that each agent can understand the environment and cooperate with each other. Therefore, they can evaluate and make their decisions based on those of other agents.

2) ONLINE TESTING

During online testing, the king agent only needs to obtain the frequency deviation, ACE and the integral value of two while other agents only need to communicate with their superior agents to obtain power regulation command, so as to realize centralized- decentralized autonomous control of AGC.

D. REGULATION MILEAGE PAYMENT

According to the rules of china southern power grid (CSG), the regulation mileage of each AGC regulation unit response to the AGC generation power command is shown in formula (1) [32].

The formula for regulation mileage of frequency regulation unit i is as follows:

$$M_i(k) = \left|\Delta P_i^{\text{out}}(k+1) - \Delta P_i^{\text{out}}(k)\right| \tag{1}$$

where $M_i(k)$ refers to the frequency regulation mileage of the ith AGC unit within the kth control interval period. In this formula, $\Delta P_i^{\text{out}}(k)$ refers to the actual regulation power output of the ith AGC unit within the period of the kth control interval.

The regulation mileage payment can be calculated by the following formula [32]:

$$D_i = \sum_{k=1}^{N} \lambda \cdot S_i^{\text{p}} \cdot M_i(k) \tag{2}$$

where D_i refers to the total regulation mileage payment of the ith AGC unit in N control intervals; λ refers to the price of the frequency regulation mileage, S_i^{p} means the comprehensive frequency regulation performance indicator score of the ith AGC unit; additionally, N refers to the control interval number within every period of frequency regulation service. For example, when the time cycle of AGC control is set to be 4s, the real-time frequency regulation market settlement cycle will be 900s. In addition, the amount of N is 225.

E. OBJECTIVE FUNCTION

1) KING AGENT

f_1 is the absolute value of total frequency deviation, and f_2 is the absolute value of the total ACE, the objective function can be expressed as formula (3)

$$\begin{cases} \min\ f_1 = \sum_{k=1}^{N} |\Delta f(k)| \\ \min\ f_2 = \sum_{k=1}^{N} |e_{ACE}(k)| \end{cases} \tag{3}$$

where n is the number of AGC units; $\Delta f(k)$ is the frequency deviation of control interval k, and $P_i^{\text{out}}(k+1)$ is output of AGC unit i at the beginning of control interval $k+1$.

2) GENERAL AGENT

In the objective function of general agent, for VGA-AGC, the frequency error and regulation mileage payment are fully considered, thus taking into account both the control performance and the frequency regulation mileage payment for optimization. The objective function is as follows:

$$\begin{cases} \min f_{G1} = \sum_{k=1}^{N} \Delta P_{\text{error-G}}^2(k) \\ \min f_{G2} = \sum_{k=1}^{N} \sum_{i=1}^{5} D_i(k) \\ \Delta P_{error-G}(t) = \left| \Delta P_{order\text{-}\sum}(k) - \Delta P_{\text{G-}\sum}(k+1) \right| \end{cases} \tag{4}$$

where $\Delta P_{\text{error-G}}(k)$ is total power control error at control intervals k, $\Delta P_{\text{order-}\sum}(k)$ is the total AGC generation power command of control interval k, and $D_i(k)$ is the regulation mileage payment of unit group i in the control interval k.

3) LORD AGENT *J*

$$\begin{cases} \min f_{L1} = \sum_{k=1}^{N} \Delta P_{\text{error-L}}^2(k) \\ \min f_{L2} = \sum_{k=1}^{N} \sum_{i=1}^{n} d_i(k) \\ \Delta P_{\text{error-L-}j}(k) = \left| \Delta P_{\text{order-L-}j}(k) - \Delta P_{\text{G-L-}j}(k+1) \right| \\ d_j^{n_j}(k) = \lambda^* S_{n_j}^{\text{p}*} \left| \Delta P_{Gj}^{n_j}(k) - \Delta P_{Gj}^{n_j}(k+1) \right| \end{cases} \tag{5}$$

where $\Delta P_{error\text{-}L}(k)$ is the power control error of control interval k, $d_i(k)$ is regulation mileage payment of unit group i in control interval k, $\Delta P_{\text{error-L-}j}(k)$ is the total power control error of territory unit group j in control interval k, and $\Delta P_{\text{order-L-}j}(k)$ is the total AGC generation power command of territory unit group j in control interval k; $\boldsymbol{\Delta P}_{\boldsymbol{Gj}}^{\boldsymbol{nj}}(\boldsymbol{k})$ is the actual total regulation output of territory unit group j at the beginning of control interval k, $\boldsymbol{S}_{\boldsymbol{nj}}^{\boldsymbol{p}}$ is the comprehensive frequency regulation performance index of unit n_j.

F. FREGULATION UNIT AND RELEVANT CONSTRAINTS

1) CONVENTIONAL REGULATION UNIT

The conventional regulation unit includes coal-fired unit. LNG unit and hydroelectric unit.

Constraints: Power balance constraints, regulation direction constraints, the upper and lower limits of AGC regulation capacity, and constraints on generation ramp rate, which are as formulas (11) and (12) in sequence:

$$\begin{cases} \sum_{i=1}^{n} \Delta P_i^{\text{in}}(k) = \Delta P_{\text{order-}\sum}(k) \\ \Delta P_{\text{order-}\sum}(k) * \Delta P_i^{\text{in}}(k) \geq 0 \\ \Delta P_i^{\min} \leq \Delta P_i^{\text{in}}(k) \leq \Delta P_i^{\max} \end{cases} \tag{6}$$

$$\left| \Delta P_i^{\text{out}}(k+1) - \Delta P_i^{\text{out}}(k) \right| \leq \Delta P_i^{\text{rate}} \tag{7}$$

where $\Delta P_{\text{order-}\sum}(k)$ refers to the total AGC power regulation command at the beginning of the kth control interval, $\Delta P_i^{\max}$ refers to the AGC regulation power upper limit of the ith AGC unit, and $\Delta P_i^{\min}$ refers to the AGC regulation power upper limit of the ith AGC unit. In addition, ΔP_i^{rate} refers to the ramp rate of the ith AGC unit.

2) REGULATION UNIT FOR THE CHP AND P2G

The CHP consists of a compressor, a combustion chamber and a steam turbine. The transformation formula is as follows:

$$P_{\text{CHP}} = \eta_{\text{e}} G_{\text{CHP}} \tag{8}$$

where P_{CHP} is the electrical power of the CHP unit, η_{e} is the generation efficiency, and G_{CHP} is the gas consumption power.

Constraints on the feasible operating range of CHP: Relevant constraints are as shown in formulas (7) and (9):

$$\begin{cases} P_i^{\min}(H_i) \leqslant P_i^0 + \Delta P_i \leqslant P_i^{\max}(H_i) \\ H_i^{\min}\left(P_i^0 + \Delta P_i\right) \leqslant H_i \leqslant H_i^{\max}\left(P_i^0 + \Delta P_i\right) \end{cases} \tag{9}$$

where H_i is heat output of CHP, $P_i^{\min}(H_i)$ is the lower limit of CHP electrical power output when the heat output is H_i. $P_i^{\max}(H_i)$ is the upper limit of CHP electrical power output when the heat output is H_i. P_i^0 is CHP's basic electrical power output. ΔP_i is CHP's additional electrical power output. $H_i^{\max}(P_i^0 + \Delta P_i)$ is the lower limit of CHP's heat output when the electrical power output is $P_i^0 + \Delta P_i$. $H_i^{\min}(P_i^0 + \Delta P_i)$ is the lower limit of CHP's heat output when the electrical power output is $P_i^0 + \Delta P_i$.

P2G is used to convert electrical energy into easy-to-transport-and-store hydrogen or natural gas through conversion. Its model can be described by formula (10). It can be regarded as a fast unit.

$$G_{\text{P2G}} = \frac{3600 \eta_{\text{P2G}}}{f_{\text{LHV}}} P_{\text{P2G}} \tag{10}$$

where G_{P2G} is natural gas production of the P2G, P_{P2G} is its electrical power consumption, η_{P2G} is the conversion efficiency, and f_{LHV} is the low heat value of the natural gas.

3) EGULATION UNIT OF RENEWABLE ENERGY

The frequency regulation unit of the renewable energy system includes photovoltaic power and wind turbine, which are controlled by power electronic equipment, and the constraints are as shown in formula (7).

It is assumed that the active wind turbine output is tracked and controlled based on the maximum power point. $P_{g,}$ is the active power output which can be calculated with the wind speed as follows:

$$P_g = \begin{cases} 0 & V_w < V_w^{\text{in}},\ V_w > V_w^{\text{out}} \\ P_W^{\text{base}} \dfrac{V_w - V_w^{\text{in}}}{V_w^{\text{base}} - V_w^{\text{in}}} & V_w^{\text{in}} \leq V_w \leq V_w^{\text{base}} \\ P_w^{\text{base}} & V_w^{\text{base}} \leq V_w \leq V_\omega^{\text{out}} \end{cases} \tag{11}$$

where V_w is the wind speed, V_w^{in} and V_w^{out} are the cut-in wind speed and cut-out wind speed respectively, V_w^{base} is the rated

wind speed of the fan, P_w^{base} is the rated output power of the fan. P_{pv} is the active photovoltaic power output. It can be calculated as follows:

$$P_{pv} = P_{pv}^{\text{base}} \left[1 + \alpha_{pv} \cdot \left(T - T_{ref}\right)\right] \cdot \frac{s_{pv}}{1000} \tag{12}$$

where P_{pv}^{base} is the rated generated power of the photovoltaic power station, α_{pv} is the photovoltaic temperature conversion power factor, T is the temperature at the current moment, T_{ref} is the reference value of temperature, S_{pv} is the illumination intensity at the current moment.

4) FLEXIBLE LOAD REGULATION UNIT

The load aggregators are introduced according to the power grid source-charge collaborative frequency control theory. For relevant equipment, there are three temperature control devices including air conditioner, refrigerator, electric water heater, energy storage and electric vehicles. The general constraints of them are shown in formula (7).

General heating/cooling air conditioner load is an important part of demand response. The mathematical expression for cooling air conditioning is as follows:

$$T_{t+1}^{in} = T_t^{in} e^{\Delta t/RC_{air}} + \left(e^{\Delta t/RC_{air}} - 1\right)\left(T_t^{out} - RP_t\Delta t\right) \tag{13}$$

- where Q_t is the indoor heat absorbed from the outdoor at time t; T_t^{out} and T_t^{in} are outdoor and indoor temperatures at time t respectively; R is the thermal resistance of the house, P_t is the heating power at time t, C_{air} is the specific heat capacity of air, and Δt is time increment.

The electrical characteristics of the refrigerator load are shown in formula (14):

$$T_{t+1}^{\text{FR}} = T_t^{\text{FR}} - \left(\alpha_{\text{FR}} s_t^{\text{FR}} - \gamma_{\text{FR}}\right)\Delta t \tag{14}$$

where T_t^{FR} is the internal temperature of the refrigerator at the time t; s_t^{FR} is the on/off state for refrigeration function of the refrigerator; α_{FR} is the refrigeration coefficient of the refrigerator when the refrigeration function is on; γ_{FR} is the warming coefficient of the refrigerator when the refrigeration function is off.

It is assumed that after the hot water is consumed, an equal amount of cold water will be immediately introduced. According to the second law of thermodynamics, the water temperature can be expressed as follows:

$$T_{t+1}^{\text{hw}} = \frac{\rho V_t^{\text{cold}}(T^{\text{cold}} - T_t^{\text{hw}}) + \rho V^{\tan k} T_t^{\text{hw}}}{\rho V^{\tan \text{k}}} + \frac{h_t^{\text{wh}}\Delta t}{\rho V^{\tan k} C_{\text{w}}} \tag{15}$$

where T_{hw} is the hot water temperature at time t; ρ, V^{tank}, and C_{w} are water density, water tank volume and specific heat capacity of water respectively; V_{cold}^t is the volume of cold water introduced at time t; T^{cold} is the temperature of introduced cold water; h_{wh}^t is the heating power at time t.

The battery energy storage system (BESS) can help conventional regulation units maintain frequency stability due to its fast response speed and flexible control of output.

$$SOC_{im}^{\min} \le SOC_{im}(k) \le SOC_{im}^{\max} \tag{16}$$

$$SOC_{im}(k) = \begin{cases} SOC_{im}(k-1) - P_{im}^{\text{out}}(k) \\ \cdot \Delta T \cdot \eta_{\text{ch}}/E_{im} & if\ P_{im}^{\text{out}}(k) \le 0 \\ SOC_{im}(k-1) - P_{im}^{\text{out}}(k) \\ \cdot \Delta T/\left(\eta_{\text{dis}} \cdot E_{im}\right) & if\ P_{im}^{\text{out}}(k) > 0 \end{cases} \tag{17}$$

where, $\text{SOC}_{im}^{\min}$, $\text{SOC}_{im}^{\max}$ are the upper and lower limits of SOC, η_{ch} is the charging efficiency, η_{dis} is the discharge efficiency, and E_{im} is the rated capacity.

Electric vehicles can be regarded as a kind of large-scale energy storage facilities. After a single electric vehicle is connected to the charging pile, the relation between the state of charge (SOC) at time t. P is the charging/discharging power, can be expressed as SOC. $C_{\max}$ is the ratio of current battery capacity to the battery capacity:

$$SOC_t = SOC_{t-1} + \frac{\int_{t-1}^{t} Pdt}{C_{\max}} \tag{18}$$

P is the charging/discharging power. In order to ensure the safety of the battery in the process of charging/discharging, P shall be within the charging/discharging power limit of the battery:

$$P_{\text{dis. max}} \le P \le P_{\text{char. max}} \tag{19}$$

where $P_{\text{dis.max}}$ and $P_{\text{char.max}}$ are the maximum discharging power and the maximum charging power that the battery can withstand under safe operation conditions respectively.

III. MADMI-TD3

A. DEEP DETERMINISTIC POLICY GRADIENT (DDPG)

DDPG, an actor-critic framework algorithm of deep reinforcement learning, incorporates deep learning neural networks into the deterministic policy gradient (DPG):

DDPG only employs an actor network to explore the environment, which will lead to a large amount of redundancy in the information used by agent. Thus resulting in slow parameter update. It is difficult to ensure the diversity of samples and easy to fall into the local optimum.

B. MADMI-TD3 FRAMEWORK

MADMI-TD3 is a deep reinforcement learning algorithm developed from DDPG [33], [34]. In order to overcome the over-estimation of Q value [35], [36] and the low training efficiency problem [37] in DDPG, the algorithm adopted seven techniques for improvement of the stability and training efficiency.

1) CLIPPED MULTIPLE Q-LEARNING STRATEGY

MADMI-TD3 employs the clipped multiple Q-learning strategy to calculate the target value, and the formula is as follows:

$$y_t^1 = r(s_t, a_t) + \gamma \min_{i=1,2} Q_{\theta_i'}\left(s_{t+1}, \pi_{\phi_1}(s_{t+1})\right) \tag{20}$$

To reduce the cost of training, each explorer used an independent actor network and two critic networks. $\pi_{\phi 1}$ is the strategy of actor network which is updated based on $Q_{\theta 1}$ of the critic network. y_t^2 and y_t^1 are equal. $Q_{\theta' 2}$ is the values of critic network.

2) STRATEGY DELAYED UPDATING

MADMI-TD3 updates the actor network one time after the critic network is updated d times, so as to ensure that the actor network can be updated as the Q value error is low, thus improving the updating efficiency of actor network.

3) SMOOTH REGULARIZATION OF TARGET STRATEGY

The algorithm introduced a regularization method to reduce the variance of the target values, and smooth the Q-value estimate by bootstrap of similar action:

$$y_t = r(s_t, a_t) + \mathrm{E}_\varepsilon \left[Q_{\theta'} \left(s_{t+1}, \pi_{\phi'}(s_{t+1}) + \varepsilon \right) \right] \quad (21)$$

Moreover, smooth regularization is achieved by adding a random noise to the target strategy and averaging on mini-batch:

$$y_t = r(s_t, a_t) + \gamma \min_{i=1,2} Q_{\theta_t} \left(s_{t+1}, \pi_{\phi'}(s_{t+1}) + \varepsilon \right) \quad (22)$$

$$\varepsilon \sim \mathrm{clip}\left(N(0, \sigma), -c, c\right) \quad (23)$$

4) DISTRIBUTED REINFORCEMENT LEARNING FRAMEWORK BY DECENTRALIZED IMPLEMENTATION AND CENTRALIZED TRAINING

In MADMI-TD3, each agent has multiple explorers, a leader and two shared experience pools, among which the leader includes two critic networks and an actor network. Each explorer has an actor network. Also, it has own network and environment. For an exploration environment with several different explorers, first, the explorers generate the transformation experience based on their own environment and add the transformation experience to the two experience buffer pools according to the criteria. Then the leader samples and transforms the experience from the experience buffer pools according to the criteria. After this, it keeps learning. To speed up the learning process of an agent, the input of the critic network in the leader should include the observed states and actions of other agents, so that each agent can have a comprehensive understanding of the environment, thus properly evaluating the strategy, and cooperating with other agents. Finally, the actor network in the explorer periodically updates its network parameters based on the latest actor network from the learner. Since multi-agent centralized training and decentralized execution are adopted, different explorers of multiple agents are grouped into one group. The distributed training is implement with multiple explorer teams in parallel, as described in Section 1.1.2. In the centralized training, each agent gets the action that needs to be performed in the current state according to its own strategy. After all agents interact with the environment. Each agent randomly selects experience from the experience pool to train their neural network.

5) VARIABLE NOISE MODEL

In the algorithm proposed, random noise with different variances is adopted to actor networks of different explorers, so as to produce different samples. The random noise model employs Gaussian noise or ornstein-uhlenbeck (OU) noise model randomly, so as to increase the randomness and diversity of the explored samples.

$$\text{nosie} = N(0, \sigma) \quad \text{or OU} \quad (24)$$

6) CLASSIFIED EXPERIENCE REPLAY

The algorithm proposed uses the classification criteria for the mean value of immediate rewards: two completely independent experience buffer pools are used to store experience samples. When the network model is initialized, the average value of immediate reward value of all samples in the two experience buffer pools is set to 0. During training, the immediate reward value is compared with the average value of the sample data. If the immediate reward value in the experience sample is greater than r_a which is the mean of all the immediate reward values in the experience sample, store the sample in experience buffer pool 1, otherwise store it in experience buffer pool 2.

7) "WARM BOOT" OF EXPERIENCE POOL

To improve the algorithm in optimizing features, so as not to lose the direction for optimizing due to too many low-value samples learnt by the algorithm at the beginning of the training in the early stages of the training, in this article, the experience pool is designed in a way of "warm boot", that is, before the formal training, let each agent conduct "warm-up" training, so as to produce some samples and classify the samples based on the principle of classified replay. Then according to the reward value, the sample in experience pool 1 obtained by "warm-up" training is divided into two parts according to the classification criteria, which are put into experience pool 1 and experience pool 2 respectively. In both experience pools, samples of "warm-up" training with high reward value are left in advance, so that the algorithm will not to drift during formal training, thus obtaining a better solution and accelerating the convergence speed.

During training, as to pool 1, n_ξ samples can be gotten with the probability of ξ. In pool 2, $n_{(1-\xi)}$ samples can be gotten with the probability of 1-ξ. The detailed framework is displayed in Fig. 2. The explicit process is displayed in Table. 1.

IV. VGA-AGC SYSTEM BASED ON MADMI-TD3 IN PERFORMANCE-BASED FREQUENCY REGULATION MARKET

A. KING AGENT-CONTROLLER AGENT

1) ACTION SPACE

For any control interval, the output of king agent is the total generation power command $\Delta P_{\text{order-}\sum}(k)$, and so the action space is as follows:

$$\left[\Delta P_{\text{order-}\sum}(k)\right] \quad (25)$$

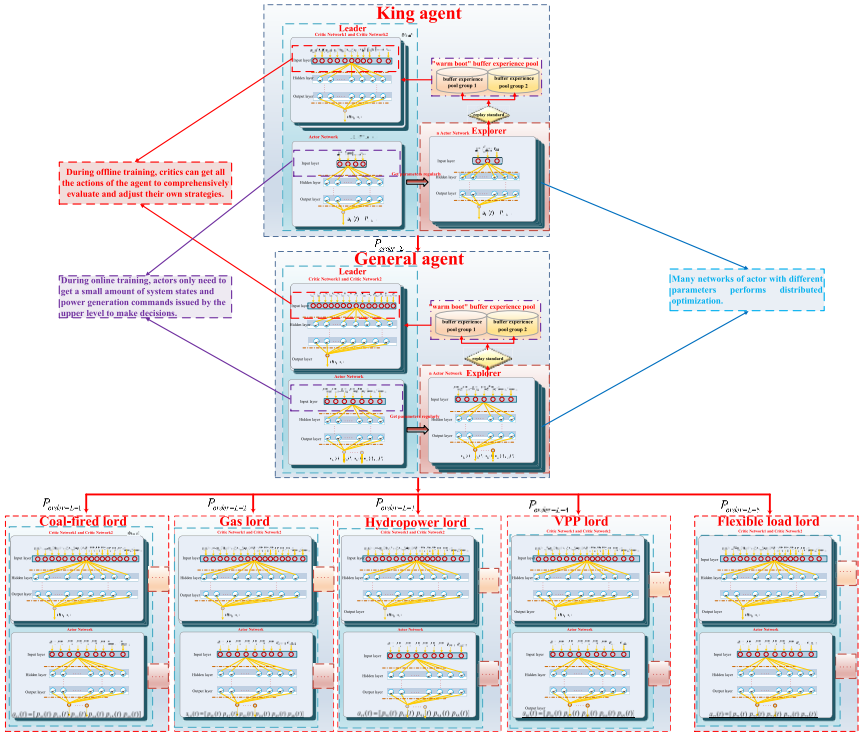

FIGURE 2. VGA-AGC framework based on MADMI-TD3 algorithm.

TABLE 1. VGA-AGC algorithm flow.

1: **INITIALIZE** parameters θ_1 and θ_2 of critic networks in learner, as well as parameter ϕ of the actor network. Initialize the target network parameter, Experience buffer pools: D_1 and D_2, discount factor γ, maximum episode number: E; maximum time step of each episode: $T_{\max}$, sampling probability for experience buffer pool 1: ξ, sampling probability for experience buffer pool 1:1-ξ. Complete the warm boot of the experience pools, group the multi-agent explorers to form explorer groups, 7 actors networks each group, with each agent corresponding to one of the actors networks.
2: **FOR** *episode* =1, E do (Train E episodes from 1 to E)
Explorer team:
3: Initialize the noise$_i$ of each actor network for explorer group i (i = 1, 2, 3,...,n).
4: Get the initial state s_i^1 of the actor network for explorer group i.
5. Get parameter ϕ_i for each actor network from leader.
6: In explorer group i, **FOR** t=1, $T_{\max}$ do
7: Select the action $a_i^t = \mu(s_i^t|\mathbf{0}_i^\mu) + noise_t$.
8: Execute the action to get the immediate reward value and the next state r_{i+1}^t.
9. Store the immediate reward values from the experience sample $e_i = (s_i^t, a_i^t, r_i^t, s_{i+1}^t)$ in Experience buffer pool D_1 or D_2.
Leader:
10: Sample a mini-batch from experience buffer pools 1 and 2. With probability of ξ,1-ξ, select n_ξ experience samples from D_1, and $n_{(1-\xi)}$ experience samples from D_2: $\left(s_j^i, a_j^i, r\left(s_j^i, a_j^i\right), s_{j+1}^i\right)$, $j = 1,\ldots,N, N = n_\xi + n_{1-\xi}$.
11: Use the target network to get the action.
12: Calculate the target value for each transfer experience in minibatch.
13: Update θ_i according to the gradient$_i$.
14: **IF** t mod d then (update the actor network every d critic network updates).
15: update ϕ according to the gradient .
16: Update the target network.
END IF
END FOR
END FOR

2) STATE SPACE

The state space is compose of frequency deviation Δf, the time integral of frequency deviation ($\int_0^t \Delta f\, dt$), the ACE e_{ACE} and the time integral of it ($\int_0^t e_{ACE} dt$). It can be shown as follows:

$$[\Delta f \quad \int_0^t \Delta f\, dt \quad \int_0^t e_{ACE} dt \; e_{ACE}] \tag{26}$$

3) REWARD FUNCTION

The reward function is as follows:

$$r(t) = -\left[\mu_1 \Delta f(k)^2 + \mu_2 |e_{ACE}(k)|\right] + A \tag{27}$$

$$A = \begin{cases} 2 & |\Delta f(t)| < 0.05 \\ 0 & |\Delta f(t)| \geq 0.05 \end{cases} \tag{28}$$

where A is the control reward item, and when the absolute value of frequency deviation is less than 0.05, it is equal to 2.

B. GENERAL AGENT – TOTAL DISPATCH AGENT

1) ACTION SPACE

In order to meet the power balance constraint, as shown in formula (29), the participation factors of each generation unit are to satisfy the following formula:

$$0 \leq a_{Gi} \leq 1, \quad \sum_{i=1}^{n} a_{Gi} = 1 \tag{29}$$

To meet the requirement of formula (29), suppose that it has n units, but participation factors only need to be allocated to $n-1$ of them in each AGC dispatch period. The participation factor for the nth group can be calculated

as below:

$$a_{Gn} = 1 - \sum_{i=1}^{n-1} a_{Gi} \tag{30}$$

In this article, generation unit group n is defined as the balanced unit group, and the unit group with large regulation capacity is selected as the balancing unit group, and the action is:

$$[a_{G1} \quad a_{G2} \quad a_{G3} \quad a_{G4}] \tag{31}$$

Constraints:

$$\begin{cases} \sum_{i=1}^{4} a_{Gi} < 1 \\ a_{G5} = 1 - \sum_{i=1}^{4} a_{Gi} \end{cases} \tag{32}$$

2) STATE SPACE

State space of the general agent is composed of the total AGC generation power command $\Delta P_{\text{order-}\sum}$, ΔP_{GG1}, ΔP_{GG2}, ΔP_{GG3}, ΔP_{GG4}, ΔP_{GG5} which are the outputs of different units respectively. It can be shown in formula (33).

$$[\Delta P_{\text{order-}\sum} \quad \Delta P_{GG1} \quad \Delta P_{GG2} \quad \Delta P_{GG3} \quad \Delta P_{GG4} \quad \Delta P_{GG5}] \tag{33}$$

3) REWARD FUNCTION

In order to keep the forms of reward function consistent, the score of regulation mileage comprehensive frequency regulation performance index is the historical average value obtained through a long-term simulation.

$$r(k) = -\left[\mu_1^G \Delta P_{error-G}^2(k) + \mu_2^G \sum_{i=1}^{5} D_i(k)\right] + \sigma \tag{34}$$

$$\sigma = \begin{cases} -0.5 & a_{Gn} < 0 \\ 0 & a_{Gn} \geq 0 \end{cases} \tag{35}$$

where σ is the penalty term. If the participation factor of balanced units is less than 0 σ is equal to -0.5.

C. LORD AGENT J – SUB DISPATCH AGENT

Several lord agents have similar state space, action space and reward functions.

1) ACTION SPACE

In this article, unit n is defined as a balanced unit, For the lord agent $j(j = 1, 2, 3, 4, 5)$, the action space is as follows:

$$\begin{cases} \left[a_{Lj1}\ a_{Lj2}\ \ldots a_{Lji} \ldots a_{Lj(n-1)}\right], & \sum_{i=1}^{n-1} a_{Lji} < 1 \\ a_{Ljn} = 1 - \sum_{i=1}^{n-1} a_{Lji} \end{cases} \tag{36}$$

2) STATE SPACE

State space of the general agent is composed of the AGC generation power command $\Delta P_{\text{order-}j}$ input by the lord agent j and the actual output of n units $\Delta P_{Gi}(i = 1, 2, 3, \ldots, n)$, which are managed by the lord agent j. It can be shown in formula (37).

$$[\Delta P_{order-i} \quad \Delta P_{G1\text{-}L\text{-}j} \quad \Delta P_{G2\text{-}L\text{-}j} \quad \ldots. \Delta P_{Gn\text{-}L\text{-}j}] \tag{37}$$

3) REWARD FUNCTIONS

As shown in formulas (38)-(39).

$$r(k) = -\left[\mu_1^L \Delta P_{error\text{-}L\text{-}j}^2(t) + \mu_2^L \sum_{n_j=1}^{n_j} d_{L-j}^{n_j}(t)\right] + \rho \tag{38}$$

$$\rho = \begin{cases} -0.5 & a_{Ljn} < 0 \\ 0 & a_{Ljn} \geq 0 \end{cases} \tag{39}$$

D. PARAMETER SELECTION

The weight coefficient in the reward function and the hyper-parameter design in the pre-learning are as shown in Table 2.

TABLE 2. Parameter settings.

Parameter	Value
Critic learning rate	0.0009
Actor learning rate	0.0009
Discount factor	0.9
Number of explorer groups	5
Selection probability for Experience pool 1	0.89
Target action noise variance	0.015
Updating interval of strategy network	2
Size of Experience pools 1,2	3000000
Noise model of Explorer groups 1~3	gaussian noise
Noise model of Explorer groups 4~5	OU noise

V. SIMULATION VERIFICATION

To verify the effectiveness of the proposed VGA-AGC based on MADMI-TD3, the conventional AGC framework (PI + GA, PI + PROP) and VGA-AGC based on deep reinforcement learning algorithm (MADMI-DDPG, TD3 and DDPG) are introduced as the comparisons.

The interconnected power grid system of a province includes 32 regulation units. The specific control model is shown in Figure 3, and the parameters for the units are shown in Table 5.

A. SIMULATION OF A PROVINCIAL INTERCONNECTED POWER GRID UNDER RANDOM STEP DISTURBANCE

1) PRE-LEARNING STAGE

In the pre-learning stage, a durative sinusoidal disturbance with cycle of 3600s, amplitude of 1800MW, duration of

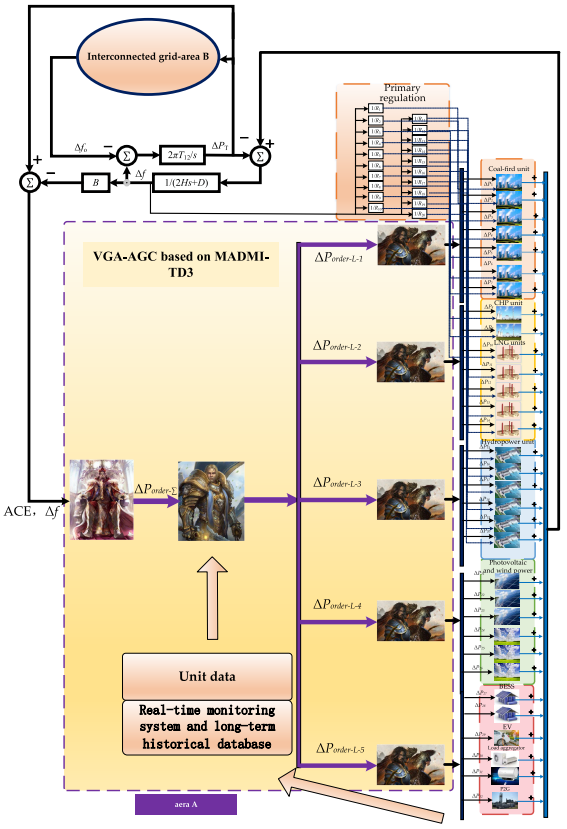

FIGURE 3. A provincial power grid AGC control chart.

3600s and phase of 0.5π is added to area A. The specific control model diagram is shown in Figure 3, and the parameters for the generation units are shown in Table 2. The training chart is shown in Figure 4.

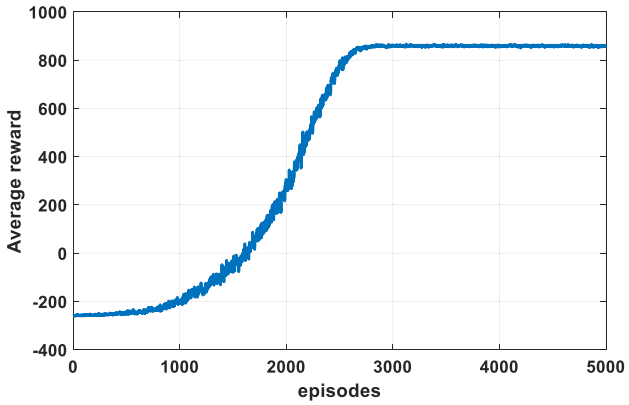

FIGURE 4. Training chart.

In Fig. 4, the curve represents the mean of reward values of corresponding episodes for MADMI-TD3. Obviously, The average reward value of the MADMI-TD3 algorithm can smoothly converge to an optimal solution, and the algorithm is stable.

2) STEP DISTURBANCE ONLINE TEST

For a power grid containing various regulation units, step load disturbance is used for testing. The amplitude is not more than 800MW. The results are shown in Figures 5-9 and Table 4.

Figures 5-9 provide the online test results of VGA-AGC based on MADMI-TD3 MADMI-DDPG, MA-TD3 and MA-DDPG as well as the conventional combined algorithm (PI + GA, PI + PROP). According to Fig. 5, at the same

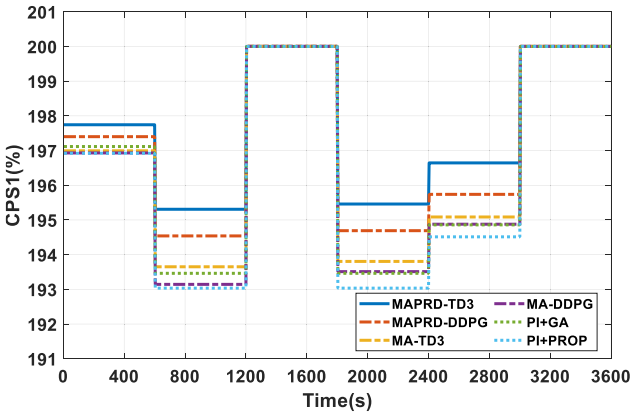

FIGURE 5. CPS1 in 10min of a provincial grid with step disturbance.

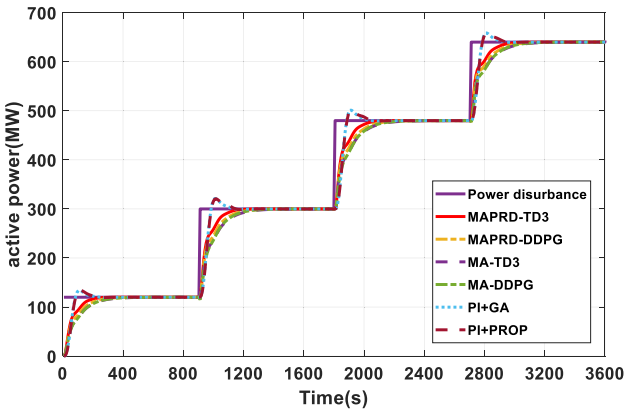

FIGURE 6. Total power regulation of a provincial grid with step disturbance.

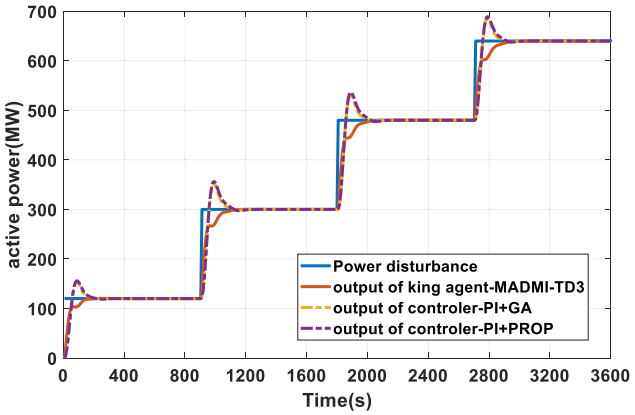

FIGURE 7. Total power generation command of a provincial grid with step disturbance.

TABLE 3. Results of a provincial grid with step disturbance.

Algorithm	$\|\Delta f\|$/Hz	$\|e_{ACE}\|$/MW	C_{CPS1}/%	Payment/$
MADMI-TD3	0.0490	9.861	197.531	646.785
MADMI-DDPG	0.0562	12.661	197.068	747.072
MA-TD3	0.0637	14.772	196.598	733.196
MA-DDPG	0.0647	14.772	196.419	747.072
PI+GA	0.0734	10.447	196.316	795.932
PI+PROP	0.0648	10.084	196.259	805.458

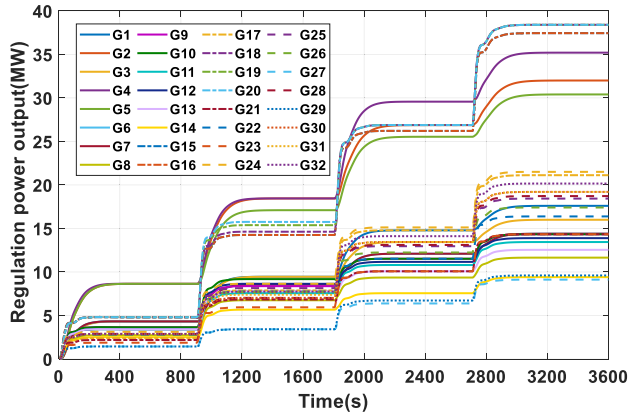

FIGURE 8. Unit regulation power output of a provincial grid with step disturbance.

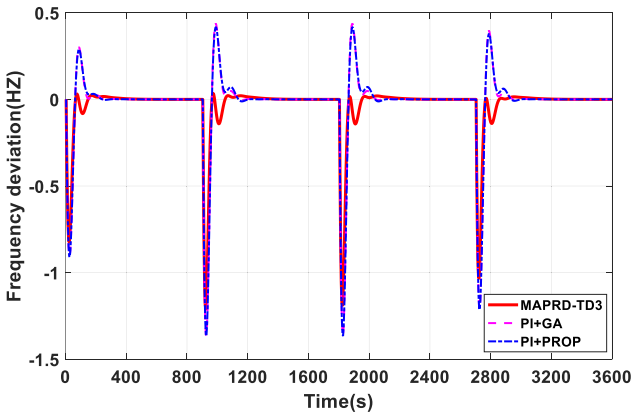

FIGURE 9. Frequency deviation of a provincial grid with step disturbance.

moments, the 10-minute average CPS1 of MADMI-TD3 is significantly higher than that of the other six algorithms. MADMI-TD3 power control deviation is much smaller than that of conventional combination algorithm and the response rate is faster than that of conventional combination algorithm according to Fig. 6. This shows that the actual total output of generation unit by MADMI-TD3 algorithm is closer to the actual load disturbance. The reason for this is that in conventional combination algorithm, too many slow-response units are used for frequency regulation. In addition, the coordination of PI controller and generation command dispatch algorithm is not taken into account, which may cause fluctuation of total AGC generation power command output by PI controller (as shown in Fig. 7). Thus, the output of some generation units are regulated frequently, and ''overregulation'' is occurred, which can greatly increase the regulation mileage of some units, thus increasing the regulation mileage payment. In contrast, in the results of MADMI-TD3, more fast-response units are used for frequency regulation, such as hydroelectric units, renewable energy units, and flexible load (as shown in Fig. 8). Moreover, in the process of offline training, considering the coordination of agents, for king agent, there is no instability and discordance caused by the problem of cooperation between control algorithm and dispatch algorithm in the VGA-AGC based on MADMI-TD3. The king agent (controller) output can track load disturbance in real time and accurately. By MADMI-TD3, the actual total output of generation units is always close to the actual load disturbance as the control performance of AGC is significantly improved. Thus, the possibility of ''overregulation'', the regulation mileage as well as the regulation mileage payment is reduced. It also leads to a smaller change in the frequency deviation of MADMI-TD3 frequency compared with the conventional combination algorithm (as shown in Fig. 9) and the frequency recovered faster. As shown in Fig. 9, the regulation mileage payment for MADMI-TD3 algorithm is less than that for conventional combination algorithm. As shown in Table 3 for statistical results, in comparison of several algorithms, the $|\Delta f|$, $|e_{ACE}|$ and regulation mileage payment for MADMI-TD3 are at minimum while C_{CPS1} the mean CPS1 is at maximum.

Compared with other algorithms for the VGA-AGC framework, the MADMI-TD3 algorithm has the optimal control performance and the regulation mileage payment is lower than other algorithms.

B. RANDOM POWER DISTURBANCE ONLINE TEST

In the two-area power grid systems, the disturbance of photovoltaic units and wind units are simplified as random output models, which are treated as random load disturbance of AGC system. Meanwhile, part of the capacity of wind turbine and photovoltaic units participated in secondary frequency

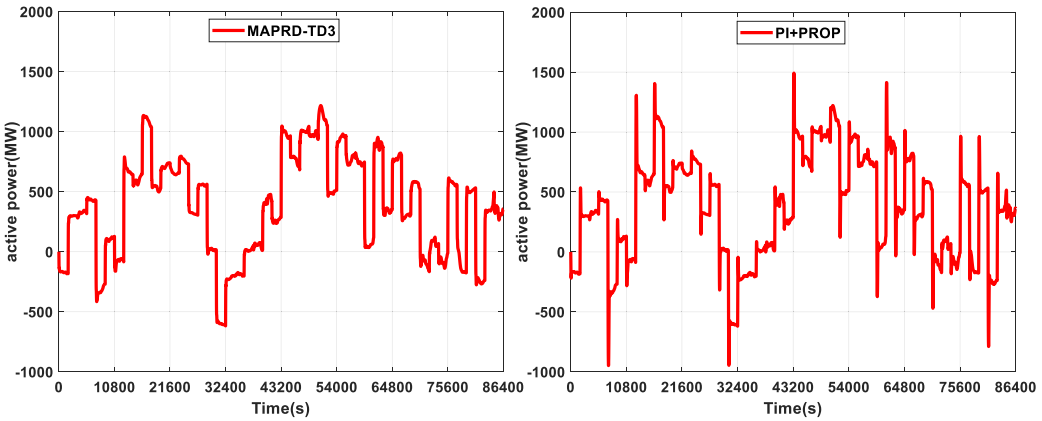

FIGURE 10. Total power generation command of a provincial grid with random disturbance.

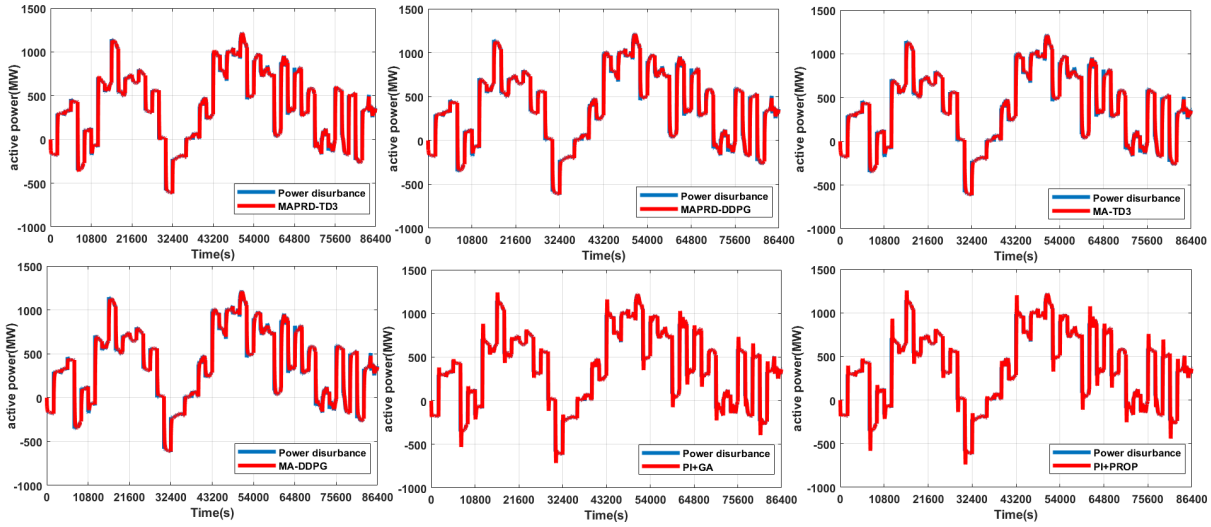

FIGURE 11. Total power regulation of a provincial grid with random disturbance.

TABLE 4. Results of a provincial grid with random disturbance.

Algorithm	$\|\Delta f\|$Hz	$\|e_{ACE}\|$/MW	C_{CPS1}/%	Payment/$
MADMI-TD3	0.0804	15.45	189.11	24071
MADMI-DDPG	0.0838	18.36	188.44	24515
MA-TD3	0.0942	21.089	187.09	24306
MA-DDPG	0.0959	21.32	186.53	24071
PI+GA	0.1116	18.00	184.05	32425
PI+PROP	0.1144	18.64	183.42	32923

regulation of the system. The wind model consists of three small capacity wind turbine units and one large offshore wind turbine unit. The latter does not participate in secondary frequency regulation. Fig. 13 shows the 24h curves of load disturbance, wind turbine's output as well as photovoltaic unit's output.

The online test results are shown in Figures 10 to 12. For the reason that there is an offshore wind power plant

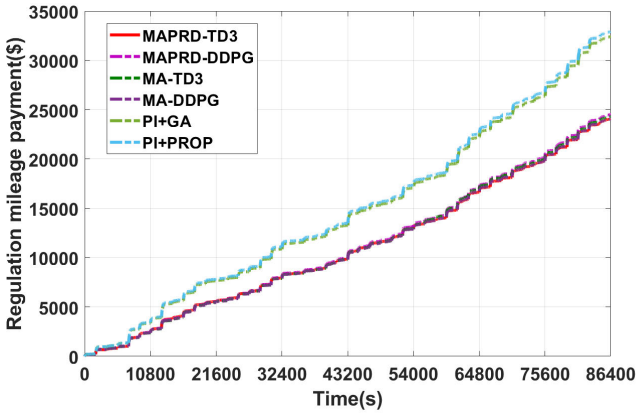

FIGURE 12. Regulation mileage payment of a provincial grid with random disturbance.

in the system, the disturbance changes very rapidly with an amplitude larger than that of the normal step disturbance. According to Fig. 10, the total generation commend output of MADMI-TD3 algorithm's king agent (controller) can still remain smooth and close to the actual power disturbance, but those of PI + PROP algorithm's controller will obviously exceed the actual load disturbance, thus appearing the overregulation of the total actual unit output and resulting in bigger frequency deviation and larger ACE, as shown in Table 4. It indicates that as the coordination of control algorithm and dispatch algorithm is considered. The MADMI-TD3 has better control performance and robustness in the control process. Fig. 11 shows the total actual regulation power output of six algorithms. Obviously, for the conventional combination algorithm, burrs and overregulation phenomenon appear more frequently. The total actual regulation power output of several deep reinforcement learning algorithm is closer to the actual load disturbance, thus obtaining better control performance and economic profits, as shown in Table 4. Because of the conventional combination algorithms have relatively large overregulation of the total actual regulation power output, the regulation mileage as well as the regulation mileage payment is increased. Hence, their regulation mileage payment is higher than that of deep reinforcement learning algorithm in each market settlement cycle (as shown in Fig. 12), indicating that as collaboration of control algorithm and dispatch algorithm is considered. Deep reinforcement learning algorithm is more economic in frequency regulation.

The results of simulation for all of the above algorithms are summarized to form Table 4. According to Table 4, the $|\Delta f\ |$, $|e_{ACE}|$ and regulation mileage payment for MADMI-TD3 algorithm are at minimum while $C_{CPS1,}$ the mean CPS1, is at maximum compare with other algorithms. The data in the Table 4 show that in the LFC which is added by random disturbance, the MADMI-TD3 has a better control effect than that of other conventional combination algorithms, and its frequency deviation is less than that of the conventional combination algorithms. However, the exploration and optimization process of other deep reinforcement learning algorithms is not optimized sufficiently. Hence, others are weaker than MADMI-TD3 algorithm in control performance and economic benefits.

VI. CONCLUSION

To conclude:

1) In the performance-based frequency regulation market, the VGA-AGC that is based on the proposed MADMI-TD3 can help build a concentrated-decentralized autonomous framework in order to solve the problem concerning the

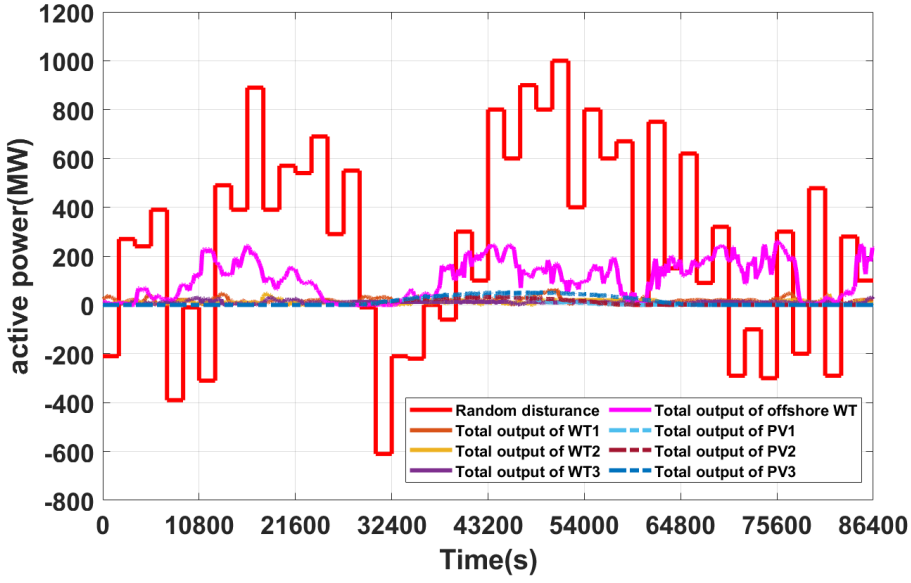

FIGURE 13. Disturbance curve.

TABLE 5. A certain province power grid AGC unit parameters.

Unit Category	No.	Unit name	T_s/s	Capacity Upper Limit/MW	Capacity Lower Limit/MW	Ramp Rate/(MW·min^{-1})
Coal-fired unit	G1	Coal-fired 1	45	330	-330	5.52
	G2	Coal-fired 2	60	180	-180	16.5
	G3	Coal-fired 3	50	240	-240	19.2498
	G4	Coal-fired 4	50	240	-240	19.2498
	G5	Coal-fired 5	60	180	-180	16.5
	G6	Coal-fired 6	42	90	-90	9
	G7	Coal-fired 7	42	90	-90	9
Gas unit	G8	CHP 1	12	200	-200	18
	G9	CHP 2	8	220	-220	15.6
	G10	LNG 1	8	220	-220	15.6
	G11	LNG 2	10	150	-150	7.2
	G12	LNG 3	10	150	-150	7.2
	G13	LNG 4	10	150	-150	7.2
	G14	LNG 5	8	220	-220	15.6
Hydropower unit	G15	Hydropower 1	5	240	-240	-
	G16	Hydropower 2	5	240	-240	-
	G17	Hydropower 3	7	160	-160	-
	G18	Hydropower 4	7	160	-160	-
	G19	Hydropower 5	8	150	-150	-
	G20	Hydropower 6	8	150	-150	-
VVP	G21	PV 1	-	40	-40	-
	G22	PV 2	-	40	-40	-
	G23	PV 3	-	40	-40	-
	G24	WT 1	-	30	-30	-
	G25	WT 2	-	30	-30	-
	G26	WT 3	-	30	-30	-
LA	G27	BESS 1	-	25	-25	-
	G28	BESS 2	-	30	-30	-
	G29	EV	-	50	-50	-
	G30	LA 1	-	30	-30	-
	G31	LA 2	-	40	-40	-
	G32	LA 3	-	50	-50	-

collaboration of conventional control algorithm and dispatch algorithm. Compared to the conventional combination algorithm, the VGA-AGC framework can realize the comprehensive optimization of control performance and economic benefits in the process of secondary frequency regulation of a power grid with large random disturbance.

2) The proposed MADMI-TD3 employs different parameters of multiple actor networks for distributed optimizing. A few techniques such as experience replay, various noise models, and warm boot of experience pool are to improve the global search ability and optimizing speed of the algorithm. The algorithm can be used to obtain the optimal control strategy in strong random environment, which can solve the problems of random disturbance caused by the large-scale distributed energies in the power grid.

3) According to the results of simulation, the proposed method can significantly improve the control performance and reduce the regulation mileage payment. Consequently, the method can obtain the maximum CPS1 index and effectively reduce the regulation mileage payment.

APPENDIX

See Table 5 and Fig. 13.

REFERENCES

[1] X. S. Zhang, T. Yu, Z. N. Pan, B. Yang, and T. Bao, "Lifelong learning for complementary generation control of interconnected power grids with high-penetration renewables and EVs," *IEEE Trans. Power Syst.*, vol. 33, no. 4, pp. 4097–4110, Jul. 2018.

[2] D. Xu, Q. Wu, B. Zhou, C. Li, L. Bai, and S. Huang, "Distributed multi-energy operation of coupled electricity, heating, and natural gas networks," *IEEE Trans. Sustain. Energy*, vol. 11, no. 4, pp. 2457–2469, Oct. 2020.

[3] L. Zhao, Z. Luo, Z. Fan, and Y. Shi, "A dual half-bridge converter with hybrid rectifier for DC power supply in railway systems," *IEEE Trans. Power Electron.*, vol. 35, no. 5, pp. 4579–4587, May 2020.

[4] H. Wang, Z. Lei, X. Zhang, J. Peng, and H. Jiang, "Multiobjective reinforcement learning-based intelligent approach for optimization of activation rules in automatic generation control," *IEEE Access*, vol. 7, pp. 17480–17492, 2019.

[5] *Technical Report on the Events of 9 Aug. 2019*. Accessed: Sep. 7, 2019. [Online]. Available: https://www.nationalgrideso.com/document/152346/download

[6] (2016). *Preliminary Report-Black System Event in South Australia on 28 Sep. 2016*. Accessed: Sep. 7, 2019. [Online]. Available: https://www.parliament.qld.gov.au/Documents/TableOffice/TabledPapers/2016/5516T1862.pdf

[7] B. K. Sahu, S. Pati, P. K. Mohanty, and S. Panda, "Teaching–learning based optimization algorithm based fuzzy-PID controller for automatic generation control of multi-area power system," *Appl. Soft Comput.*, vol. 27, pp. 240–249, Feb. 2015.

[8] R. K. Sahu, S. Panda, and N. K. Yegireddy, "A novel hybrid DEPS optimized fuzzy PI/PID controller for load frequency control of multi-area interconnected power systems," *J. Process Control*, vol. 24, no. 10, pp. 1596–1608, Oct. 2014.

[9] P. Dahiya, V. Sharma, and R. Naresh, "Automatic generation control using disrupted oppositional based gravitational search algorithm optimised sliding mode controller under deregulated environment," *IET Gener., Transmiss. Distrib.*, vol. 10, no. 16, pp. 3995–4005, Dec. 2016.
[10] F. Liu, Y. Li, Y. Cao, J. She, and M. Wu, "A two-layer active disturbance rejection controller design for load frequency control of interconnected power system," *IEEE Trans. Power Syst.*, vol. 31, no. 4, pp. 3320–3321, Jul. 2016.
[11] S. Debbarma, L. C. Saikia, and N. Sinha, "Automatic generation control using two degree of freedom fractional order PID controller," *Int. J. Electr. Power Energy Syst.*, vol. 58, pp. 120–129, Jun. 2014.
[12] T. Yu, B. Zhou, K. W. Chan, and E. Lu, "Stochastic optimal CPS relaxed control methodology for interconnected power systems using Q-Learning method," *J. Energy Eng.*, vol. 137, no. 3, pp. 116–129, Sep. 2011.
[13] T. Yu, B. Zhou, K. W. Chan, L. Chen, and B. Yang, "Stochastic optimal relaxed automatic generation control in Non-Markov environment based on Multi-Step $Q(\lambda)$ learning," *IEEE Trans. Power Syst.*, vol. 26, no. 3, pp. 1272–1282, Aug. 2011.
[14] T. Yu, H. Z. Wang, B. Zhou, K. W. Chan, and J. Tang, "Multi-agent correlated equilibrium $Q(\lambda)$ learning for coordinated smart generation control of interconnected power grids," *IEEE Trans. Power Syst.*, vol. 30, no. 4, pp. 1669–1679, Jul. 2015.
[15] T. Yu, B. Zhou, K. W. Chan, Y. Yuan, B. Yang, and Q. H. Wu, "R(λ) imitation learning for automatic generation control of interconnected power grids," *Automatica*, vol. 48, no. 9, pp. 2130–2136, Sep. 2012.
[16] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, Mar. 2014.
[17] L. Xi, L. Zhou, L. Liu, D. Duan, Y. Xu, L. Yang, and S. Wang, "A deep reinforcement learning algorithm for the order optimization allocation of total power in the interconnected power grids," *CSEE J. Power Energy Syst.*, vol. 6, no. 3, p. 712, Sep. 2020.
[18] L. Xi, Z. Zhang, B. Yang, L. Huang, and T. Yu, "Wolf pack hunting strategy for automatic generation control of an islanding smart distribution network," *Energy Convers. Manage.*, vol. 122, pp. 10–24, Aug. 2016.
[19] S. Mirjalili, "How effective is the grey wolf optimizer in training multi-layer perceptrons," *Int. J. Speech Technol.*, vol. 43, no. 1, pp. 150–161, Jul. 2015.
[20] S. Saremi, S. Z. Mirjalili, and S. M. Mirjalili, "Evolutionary population dynamics and grey wolf optimizer," *Neural Comput. Appl.*, vol. 26, no. 5, pp. 1257–1263, Dec. 2014.
[21] S. Bahrami, R.-A. Hooshmand, and M. Parastegari, "Short term electric load forecasting by wavelet transform and grey model improved by PSO (particle swarm optimization) algorithm," *Energy*, vol. 72, pp. 434–442, Aug. 2014.
[22] S. Mirjalili, "Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm," *Knowl.-Based Syst.*, vol. 89, no. 3, pp. 228–249, Nov. 2015.
[23] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Adv. Eng. Softw.*, vol. 95, pp. 51–67, May 2016.
[24] S. Mirjalili, "The ant lion optimizer," *Adv. Eng. Softw.*, vol. 83, pp. 80–98, May 2015.
[25] S. Mirjalili, "Dragonfly algorithm: A new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems," *Neural Comput. Appl.*, vol. 27, no. 4, pp. 1053–1073, May 2015.
[26] S. He, Q. H. Wu, and J. R. Saunders, *A Group Search Optimizer for Neural Network Training*. Berlin, Germany: Springer, 2006, pp. 934–943.
[27] X. Meng, Y. Liu, X. Gao, and H. Zhang. *A New Bio-Inspired Algorithm: Chicken Swarm Optimization*. Cham, Switzerland: Springer, 2014, Art. no. 103330.
[28] S. Mirjalili, "SCA: A sine cosine algorithm for solving optimization problems," *Knowl.-Based Syst.*, vol. 96, pp. 120–133, Mar. 2016.
[29] X. Meng, Y. Liu, and X. Gao, *A New Bio-Inspired Algorithm: Chicken Swarm Optimization*. Hefei, China: Springer, 2014.
[30] X. Zhang, Z. Xu, T. Yu, B. Yang, and H. Wang, "Optimal mileage based AGC dispatch of a GenCo," *IEEE Trans. Power Syst.*, vol. 35, no. 4, pp. 2516–2526, Jul. 2020.
[31] M. Clerc, "TRIBES a parameter free particle swarm optimizer," presented at the OEP, Paris, France, Apr. 2, 2003.
[32] X. Zhang, T. Tan, B. Zhou, T. Yu, B. Yang, and X. Huang, "Adaptive distributed auction-based algorithm for optimal mileage based AGC dispatch with high participation of renewable energy," *Int. J. Electr. Power Energy Syst.*, vol. 124, Jan. 2021, Art. no. 106371.
[33] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, "A review of deep learning for renewable energy forecasting," *Energy Convers. Manage.*, vol. 198, Oct. 2019, Art. no. 111799.
[34] H. Wang, Y. Liu, B. Zhou, C. Li, G. Cao, N. Voropai, and E. Barakhtenko, "Taxonomy research of artificial intelligence for deterministic solar power forecasting," *Energy Convers. Manage.*, vol. 214, Jun. 2020, Art. no. 112909.
[35] H. Z. Wang, G. B. Wang, G. Q. Li, J. C. Peng, and Y. T. Liu, "Deep belief network based deterministic and probabilistic wind speed forecasting approach," *Appl. Energy*, vol. 182, pp. 80–93, Nov. 2016.
[36] C. Vandoros, M. Geitona, V. Kontozamanis, and A. Karokis, "PHP21 pharmaceutical policy in Greece: Recent developments and the role of pharmacoeconomics," *Value Health*, vol. 8, no. 6, p. A187, Nov. 2005.
[37] 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," 2016, *arXiv:1509.02971*. [Online]. Available: https://arxiv.org/abs/1509.02971

JIAWEN LI received the M.S. degree in electrical engineering from Northeast Electric Power University, Jilin City, China, in 2016. He is currently pursuing the D.Eng. degree in electrical engineering with the School of Electric Power, South China University of Technology. His research interest includes automatic generation control.

TAO YU is currently a Professor of Power System with the School of Electric Power, South China University of Technology (SCUT), Guangzhou, China, where he is also a Professor. His research interest includes nonlinear and coordinated control theory.