Deep Reinforcement Learning Based Multi-Objective Integrated Automatic Generation Control for Multiple Continuous Power Disturbances

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 paper proposes a multi-objective integrated automatic generation control (MOI-AGC) that combines a controller with a dispatch together. This can contribute to improving both control performance and economy in a power grid with multiple continuous power disturbances. Subsequently, a distributed classification replay twin delayed deep deterministic policy gradient (DCR-TD3) is designed for MOI-AGC. On the one hand, DCR-TD3 introduces the classification replay method based on multiple explorers with different parameter actor networks for distributed optimization. On the other hand, the optimal control strategy is obtained through DCR-TD3 in an extremely random environment based on frequency deviation, regional control error together with frequency mileage payment as the reward function. This helps address the problem of frequency instability caused by multiple stochastic disturbance in a grid with a large number of distributed energies. Simulation verification is performed for the two-area load frequency control (LFC) model, with the result showing that the proposed algorithm has better control performance and economic benefits. Besides, compared with the existing algorithms, it can achieve a regional optimum control, reducing frequency mileage payment.

INDEX TERMS
Performance-based frequency regulation market, multi-objective integrated automatic generation control, distributed classification replay twin delayed deep deterministic policy gradient, regulation mileage payment, multiple continuous power disturbances.

I. INTRODUCTION
The continuous increase of the distributed energy utilization rate in the power grid is followed by the constant increase of stochastic disturbance. Thus, traditional AGC methods cannot meet such a large-scale complicated energy system’s requirement for frequency stability anymore [1].

Generally, traditional AGC generates and dispatches the system’s total AGC power regulation command to each generation unit with various methods, such as proportion integration (PI) controller, proportional dispatch based on simple generator unit capacity, and regulation speed. However, it is difficult to meet the requirement of power grid control performance standard (CPS) appraisal in the control area with a high renewable energy penetration rate and insufficient regulated sources [2].

To address the secondary frequency regulation in the power grid involving multiple renewable energy, scholars classified algorithms as two categories. One is the control algorithm for AGC, such as the proportion integration differentiation (PID) commonly employed in engineering, the fuzzy PID, fuzzy logic control [2]–[4], sliding mode control (SMC) [5], active disturbance rejection control (ADRC) [6], fractional order PID (FOPID) [7], the Q-learning in reinforcement learning, the DQN(deep Q-learning) [8], [9] learning and others. Usually, such algorithms treat the power grid as a single area to calculate the total AGC power regulation command, and then dispatch it to each AGC unit based on proportion (PROP). The other category is optimized algorithms for AGC dispatch, such as genetic algorithm (GA), gray wolf optimizer (GWO) [10], [11], proportion (PROP) [12], particle swarm optimization (PSO) [13], moth-flame optimization (MFO) [14], whale optimization algorithm (WOA) [15], ant lion optimizer (ALO) [16], dragonfly algorithm (DA) [17],
group search optimizer (GSO) [18], chicken swarm optimization (CSO) [19], sine cosine algorithm (SCA) [20] etc. Generally, such algorithms adopt the classic PID control algorithm as the control algorithm for AGC, then dispatch the total AGC power regulation command to each AGC unit based on the optimized algorithms, aiming to minimize the regulation payment. Overall, these two categories of algorithms have certain advantages. For example, control algorithms and optimized algorithms are separated, making it possible to design different algorithms separately, but there is a problem of collaboration between the two algorithms. The control algorithm takes minimizing control frequency deviation as its control objective, while the optimized dispatch algorithm takes minimizing the regulation payment as its optimization objective. Furthermore, the combination of the two algorithms will increase both the frequency deviation and the regulation payment, thus worsening the generation control performance. The actual calculation time, due to precision improvement, will exceed the maximal time allowed by the regulation command [21].

To encourage faster response AGC frequency regulation unit to participate in secondary frequency regulation, the Federal Energy Regulatory Commission (FERC) issued the No. 755 order, proposing a more fair and reasonable mechanism for the performance-based frequency regulation market (hereinafter referred to as the frequency regulation market) [22]. Based on this mechanism, the payment of each AGC unit consists of two parts [23], namely, AGC capacity payment and regulation mileage payment, which are directly affected by the frequency regulation mileage quotation. That is, the AGC frequency regulation unit with a higher quotation will be more involved in AGC frequency regulation for a better control performance. Some independent system operators (ISO), such as PJM, China Southern Power Grid (CSG) and others, invoke frequency regulation resources in accordance with the actual frequency regulation performance of different AGC units. Guided by a new frequency regulation market mechanism, the regulation payment has changed from the original simple linear calculation method with a fixed price for each unit regulation power capacity to a dynamic calculation method with the regulation mileage payment affected by the comprehensive regulation performance score, regulation mileage and the regulation mileage quotation. This makes the original method of combining the generation control algorithm with the optimization dispatch algorithm no longer suitable for the current frequency regulation mechanism. In a power grid, multiple continuous power disturbances appear, causing the system frequency to show a trend of deterioration. Even more, the mutual influence and conflict between the control performance of the generation control algorithm and the optimization performance of the optimization dispatch algorithm are deteriorated. In [24], the total AGC power regulation command is generated by the traditional PI controller, followed by the adoption of a fixed dispatch method: the participation degree of all AGC regulation units is proportional to their regulation mileage capacity [24]. Besides, in the research [25], a real-time AGC scheduling method [25] is designed, making the fast regulation resources assume a relatively huge AGC generation power command required by a high regulation mileage. A multi-agent deep reinforcement learning algorithm with action exploration and recognition thinking was proposed in [26], effectively alleviating frequency instability caused by multiple stochastic disturbance in a grid with numerous distributed energies. However, its controller only outputs total AGC power regulation command, without in-depth research on the dispatch algorithm. This tends to cause poor performance due to the uncoordinated collaboration between the control algorithm and the optimization algorithm. To address the poor coordination brought by the combination of the generation control algorithm and the optimization algorithm in microgrid, an integrated framework was designed with an adaptive deep dynamic programming [27]. This framework can prevent the poor performance brought by the combination of the control algorithm and the optimization algorithm. However, introducing the economic scheduling problem into AGC, it is not suitable for large power grid with a large number of units. In addition, the regulation payment under the frequency regulation market is ignored.

These methods, though simple and suitable, ignore the coordination of the two algorithms and specific optimization. Therefore, they cannot meet the ISO’s requirement for an optimal comprehensive benefit of control performance and regulation mileage payment.

To overcome the weakness of the combination of the generation control algorithm and the optimization dispatch algorithm in the frequency regulation market, MOI-AGC framework is designed, and DCR-TD3 is proposed for this framework. This algorithm uses multiple actor networks with different noise parameters for distributed optimum-seeking and finds the optimal strategy in an extremely random environment-interconnection power grids with large-scale distributed energy by using frequency deviation, the area control error (ACE) and the regulation mileage payment as the comprehensive reward function.

The innovation of this paper is as follows:

1) Previous studies of AGC ignored the coordination of control algorithm and optimization algorithm, thus not satisfying the benefit pursuit of ISO. It is especially true of multi-objective optimization in a performance-based frequency regulation market of power grid with large-scale distributed energy. To fill up this gap, a MOI-AGC is proposed to balance the technical and economic benefits of ISO when it distributes the total power regulation command to all the regulation units.

2) The conventional reinforcement learning continuously updates a single neural network of an agent with a single experience pool during parameter updating, causing an extremely slow parameter pace and a local optimum tendency. It is especially true of the Deep Deterministic policy gradient (DDPG) which requires high exploration capability. To address this problem, DCR-TD3 is proposed...
for MOI-AGC. Employing multiple explorers with different parameter actor networks for distributed optimization, this algorithm introduces the classification replay method for the optimal control strategy in an extremely random environment.

II. DCR-TD3 ALGORITHM

A. TWIN DELAYED DEEP DETERMINISTIC POLICY GRADIENT (TD3)

TD3 [28], deriving from DDPG [29], is a deep reinforcement learning algorithm with an actor-critic framework. In order to solve the overestimation problem of the Q value in the actor-critic framework, TD3 uses three key tricks.

1) Clipped double Q learning under the actor-critic framework [28]. TD3 selects the optimal action using the current actor network and uses the evaluation strategy of the target critic network.

\[ y_t = r(s_t, a_t) + \gamma Q(\phi_{t+1}, \pi(\theta_{t+1})) \]  (1)

The target value is as follows [31]:

\[ y^1_t = r(s_t, a_t) + \gamma \min_{i=1,2} Q(\phi^i_{t+1}, \pi^i_{t+1}) \]  (2)

where \( Q_{\theta_1} \) and \( Q_{\theta_2} \) are the value functions of two current networks under the same state \( s_t \) and the action \( \pi(\theta_t) \).

To reduce the training payment, two critic networks and an independent actor network are adopted in TD3.

2) Strategy delayed updating [28]. The critic network can reduce its deviation with the target Q value through multi-step updating. However, in the event of a big critic network deviation, updating the actor network causes divergence of the strategy. Therefore, TD3 updates the actor network once after the critic network is updated \( d \) times.

3) Smooth target strategy regularization [28]. Similar to DDPG, as TD3 uses a deterministic strategy, the target value can be easily affected by the function's approximate deviation during critic updating, resulting in an inaccurate target value. Therefore, a regularization strategy is introduced to reduce the target value variance to perform Q value estimation smoothing for the estimated value by bootstrapping a similar state action in TD3.

\[ y_t = r(s_t, a_t) + E_\epsilon \left[ Q(\phi_{t+1}, \pi(\theta_{t+1}) + \epsilon) \right] \]  (3)

At the same time, smooth regularization is realized by adding a random noise \( \epsilon \) to the target strategy and obtaining an average value from the mini-batch:

\[ y_t = r(s_t, a_t) + \gamma \min_{i=1,2} Q(\phi^i_{t+1}, \pi^i_{t+1}) + \epsilon \]  (4)

\[ \epsilon \sim \text{clip}(N(0, \sigma), -c, c) \]  (5)

B. DCR-TD3 FRAMEWORK

The traditional reinforcement learning algorithm continuously updates a single neural network of a single agent during parameter updating [30]. This results in a large amount of redundancy of information used in agent updating, an extremely slow parameter pace, and a tendency of a local optimum [31]. Especially for the deep deterministic policy gradient which requires high exploration capability.

The DDPG adopts the method of adding noise [32]. However, it is difficult to ensure sample diversity to use only one actor network for environment exploration [33]. To solve this problem, the distributed reinforcement learning training method is introduced into the TD3. At the same time, classification replay is introduced to propose the DCR-TD3.

The training framework adopted by the DCR-TD3 is a distributed reinforcement learning training framework with multiple explorers, a leader and two shared experience buffer pools. The leader includes two critic networks and an actor network. Each explorer has one actor network, its own network model and environment. The explorers first generate transfer experience based on their own environment and add the transfer experience into each of the two experience buffer pools based on the corresponding standard. Afterwards, the leader performs sampling for the transfer experience from the experience buffer pools based on the corresponding standard and keeps learning. Finally, the actor networks of the explorers regularly update their network parameters from the latest leader actor network. The actor networks of different explorers use random noise with different variances for optimum-seeking strategies to increase the randomness and diversity of the explored samples. As the algorithm adopts an experience replay mechanism, the mechanism uses a classification probability replay (short as classification replay) method. Standard for the classification probability replay as fellows: Through the \( \epsilon \)-greed strategy in Q-learning, two independent experience buffer pools are used to store empirical samples in DCR-TD3. During network model initialization, the average reward of all samples in the two pools is set to 0. Compare a sample’s reward with the average value. If the sample’s immediate reward value is larger than the average value \( \omega_1 \), place the sample in pool 1. Otherwise, place the sample in pool 2.

During training, \( n_k \) samples are selected with the probability of \( \xi \) from pool 1, and \( n(1-\xi) \) samples are selected from pool 2 with the probability of 1-\( \xi \). The specific framework is shown in Fig. 1. The specific procedure is shown in Table. 1.

III. DISTRIBUTED CLASSIFICATION REPLAY TWIN DELAYED DEEP DETERMINISTIC POLICY GRADIENT

A. FREQUENCY REGULATION MILEAGE

Frequency regulation mileage is a new quantitative indicator [14] that judges the actual regulation amount of each AGC frequency regulation unit based on the AGC generation power command allocated by the power grid scheduling center in real time. Based on the rules of the CSG market, the total frequency regulation of a certain period is the sum of the regulation mileage of all AGC frequency regulation units responding to the AGC power regulation command within that period. The frequency regulation mileage of an AGC generation power unit responding to an AGC generation power command refers to the absolute value of the difference between the actual regulation power output value at the end and the regulation power output value upon the respond. The details are as follows [8]:
The frequency regulation mileage calculation formula for the $i$th unit is:

$$M_i(k) = |\Delta P_{\text{out}}^i(k+1) - \Delta P_{\text{out}}^i(k)|$$  \hspace{1cm} (6)

where $\Delta P_{\text{out}}^i(k)$ is the actual regulation power output of the $i$th AGC frequency regulation unit during the $k$th control interval and $\Delta P_{\text{out}}^i(k)$ is the actual regulation power output of the $i$th AGC frequency regulation unit during the $k$th control interval.

The regulation mileage payment of each AGC frequency regulation unit is as follows [8]:

$$D_i = \sum_{k=1}^{N} \lambda \cdot S_p^i \cdot M_i(k)$$  \hspace{1cm} (7)

where $D_i$ is the total regulation mileage payment of the $i$th AGC frequency regulation unit in $N$ control intervals, $\lambda$ is the frequency regulation mileage price, $S_p^i$ is the comprehensive frequency regulation performance indicator score of the $i$th AGC frequency regulation unit, and $N$ is the number of control intervals of each period during the frequency regulation service. If the AGC control cycle is 4s, the time cycle of the real-time frequency regulation market clearing is 15 min (900s), $N$ is 225.

$S_p^i$ is actually the frequency regulation quality indicator of the AGC frequency regulation unit and is related to the dynamic regulation characteristic of the AGC frequency regulation unit, that is, the AGC frequency regulation unit with a quicker response will receive a higher comprehensive frequency regulation performance indicator score. Based on the rules of the CSG market, $S_p^i$ consists of the following three different scores [8].

1) Ramp rate score: refers to the rate of an AGC frequency regulation unit responding to an AGC generation power command. The calculation formula is as follows:

$$S_{\text{rate}}^i = \frac{\Delta P_{\text{rate}}^i}{\Delta P_{\text{rate}}^i}$$  \hspace{1cm} (8)

where $S_{\text{rate}}^i$ is the regulation rate score of the $i$th AGC frequency regulation unit, $\Delta P_{\text{rate}}^i$ is its maximum regulation rate,
and $\Delta P_i^{\text{rate}}$ is the average regulation rate of all AGC frequency regulation units within the corresponding control area.

2) Response delay score: refers to the time delay of an AGC frequency regulation unit responding to an AGC generation power command. The score is expressed as:

$$S_i^{\text{delay}} = 1 - \frac{T_i^d}{5^{\text{min}}} \quad (9)$$

where $S_i^{\text{delay}}$ is the response time score of the $i$th AGC frequency regulation unit, and $T_i^d$ is the regulation time constant of the $i$th AGC frequency regulation unit.

3) Regulation precision score: it evaluates regulation precision based on the deviation between the allocated AGC generation power command input and the actual unit regulation power output. The score is calculated as follows:

$$S_i^{\text{pre}} = 1 - \frac{1}{N} \sum_{k=1}^{N} \left| \frac{\Delta P_i^{\text{in}}(k) - \Delta P_i^{\text{out}}(k+1)}{\Delta P_i^{\text{in}}(k)} \right| \quad (10)$$

where $S_i^{\text{pre}}$ is the regulation precision score of the $i$th AGC frequency regulation unit, $\Delta P_i^{\text{in}}(k)$ is the AGC generation power command input of the $i$th AGC frequency regulation unit at the beginning of the $k$th control interval, $\Delta P_i^{\text{out}}(k+1)$ is the unit regulation power output of the $i$th AGC frequency regulation unit at the beginning of the $(k+1)$th control interval, $\Delta P_i^{\text{in}}(k)$ is the regulation error of the $i$th AGC frequency regulation unit during each cycle, which is 1.5% of the rated regulation power output of the $i$th AGC frequency regulation unit.

Therefore, the comprehensive frequency regulation performance indicator score is the sum of the three regulation scores multiplied by their respective weight. The formulas are as follows:

$$S_i^P = \omega_1 S_i^{\text{rate}} + \omega_2 S_i^{\text{delay}} + \omega_3 S_i^{\text{pre}} \quad (11)$$

$$\omega_1 + \omega_2 + \omega_3 = 1, \quad \omega_1 \geq 0, \quad \omega_2 \geq 0, \quad \omega_3 \geq 0 \quad (12)$$

where $\omega_1$, $\omega_2$ and $\omega_3$ are the weight coefficients of different regulation scores and their values are 0.50, 0.25 and 0.25 respectively in the frequency regulation market of the CSG.

**B. MULTIPLE CONTINUOUS POWER DISTURBANCES**

The power grid is a complex nonlinear large-scale system. After a disturbance in one place is formed, it may spread to other nodes and cause disturbances in other nodes. Therefore, in the UHV (ultra-high voltage) synchronous power grid, it is impractical to consider only a single disturbance, and the disturbance should be regarded as the combined effect of multiple events in time, space, and type that cause changes in the operating state of the system. Multiple continuous power disturbances will cause drastic changes in system power loss, thereby affecting frequency stability. Therefore, AGC system with better control performance is required when encountering multiple continuous power disturbances, so that the frequency can be quickly recovered. The combination of traditional control algorithm and dispatch algorithm cannot meet this requirement.

**C. OBJECTIVE FUNCTION**

In this research, in order to achieve an optimal comprehensive benefit of control performance and economy for dynamic dispatch of AGC power regulation command, the objective is divided into the following three parts to ensure fast and accurately balance the system power disturbance. That is, the absolute value of the total frequency deviation $f_1$, the absolute value of the total control area deviation $f_2$ and the total regulation mileage payment $f_3$ are all minimal in the AGC frequency regulation process. Obviously, the first two objectives are incompatible with the third objective. They can be expressed as (13):

$$\begin{align*}
\min f_1 &= \sum_{k=1}^{N} |\Delta f(k)| \\
\min f_2 &= \sum_{k=1}^{N} |e_{ACE}(k)| \\
\min f_3 &= \sum_{i=1}^{N} D_i
\end{align*} \quad (13)$$

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

**D. CONSTRAINT CONDITION**

Constraints of coal-fired units liquefied natural gas (LNG) units, oil units and hydropower units are as follows: power balance constraint, frequency regulation direction constraint, constraint on AGC regulation capacity upper and lower limits and generation ramp rate constraint. These constraints are expressed in (14) respectively:

$$\begin{align*}
\sum_{i=1}^{n} \Delta P_i^{\text{in}}(k) &= \Delta P_{\text{order}}(k) \\
\Delta P_{\text{order}}(k) - \sum_{i=1}^{n} \Delta P_i^{\text{in}}(k) &= 0 \\
\Delta P_{\text{min}} &\leq \Delta P_i^{\text{in}}(k) \leq \Delta P_{\text{max}} \\
|\Delta P_i^{\text{out}}(k+1) - \Delta P_i^{\text{out}}(k)| &\leq \Delta P_i^{\text{rate}}
\end{align*} \quad (14)$$

where $\Delta P_{\text{order}}(k)$ is the total AGC power regulation command at the beginning of the $k$th control interval, $\Delta P_{\text{max}}$ and $\Delta P_{\text{min}}$ are the AGC regulation capacity upper and lower limits of the $i$th regulation unit respectively, and $\Delta P_i^{\text{rate}}$ is the ramp rate of the $i$th regulation unit.

**IV. MOI-AGC BASED ON DCR-TD3 FRAMEWORK**

Unlike the traditional AGC system which adopts a combination of the PI controller and a fixed dispatch algorithm, the control algorithm of the MOI-AGC based on DCR-TD3 can output AGC power regulation command for multiple AGC units at a time, which gives consideration to the comprehensive optimum of both frequency deviation and the regulation mileage payment. The system monitors calculation, stores ACE/Δf/CPS data as well as long-term historical
records of each area of the interconnected power grid and inputs frequency deviation of the interconnected power grid ($\Delta f$), area control error (ACE), and regulation power output of each unit into the MOI-AGC. The algorithm then calculates and obtains the AGC power regulation command to be sent to each AGC unit. The specific framework is shown in Fig. 2.

### A. ACTION SPACE
To ensure that the agent’s $m$ explorers can use more standard noise sequences during training, the AGC generation factors of $n$ units are used as agent actions for any moment $t$, as shown in (15).

\[
\begin{align*}
[a_1, a_2, \ldots, a_i, \ldots, a_n], & \quad a_i \leq 1 \\
\Delta P_{\text{order}-i} = a_i^* P_{\text{max}}^G & \quad \Delta f > 0 \\
\Delta P_{\text{order}-i} = a_i^* P_{\text{min}}^G & \quad \Delta f \leq 0
\end{align*}
\]

(15)

where $a_i$ is the AGC generation factor of the $i$th unit, $P_{\text{max}}^G$ is the maximal regulation capacity of the $i$th unit, and $P_{\text{min}}^G$ is the minimal regulation capacity of the $i$th unit.

### B. STATE SPACE
Input frequency deviation ($\Delta f$) of the MOI-AGC based on DCR-TD3, the integral ($\int_0^T \Delta f dt$) of time-to-frequency deviation, the area control error ($e_{\text{ACE}}$) and actual regulation power output ($\Delta P_G$) of $n$ units. There are $n + 3$ states in total, as shown in (16).

\[
[\Delta f, \int_0^T \Delta f dt, e_{\text{ACE}} \Delta P_{G1} \Delta P_{G2} \ldots \Delta P_{Gn}]
\]

(16)

### C. SELECTION OF REWARD FUNCTION
Based on (13), a comprehensive reward function is formed using the quadratic term of ACE of each control cycle, the absolute value of the area control error and the linear weight and constraint penalty term of the regulation mileage in this paper.

To ensure consistency of the reward function form, the comprehensive regulation performance metric score for regulation mileage used in this paper is the historical average value obtained from long-term simulation.

The reward function is expressed as:

\[
r(t) = -\left[ \mu_1 \Delta f(t)^2 + \mu_2 |e_{\text{ACE}}(t)| + \mu_3 \sum_{i=1}^{m} d_i(t) \right] + A
\]

(17)

\[
d_i(k) = \lambda^* S_i^p |\Delta P_{Gi}(k) - \Delta P_{Gi}(k+1)|
\]

(18)

\[
A = \begin{cases} 
10 & |\Delta f(t)| < 0.05 \\
0 & |\Delta f(t)| \geq 0.05
\end{cases}
\]

(19)

where $t$ is the discrete moment, $e_{\text{ACE}}(t)$ is the ACE of moment $t$, $\Delta f(t)$ is the frequency deviation of moment $t$, $\Delta P_{Gi}(k)$ is the actual total regulation power output of the $i$th unit at moment $t$, $\Delta P_{Gi}(t-1)$ is the actual regulation power output of the $i$th unit at moment $t-1$, $d_i(k)$ is the regulation mileage payment of the $i$th unit at moment $t$, $\lambda$ is the regulation mileage price, and $S_i^p$ is the comprehensive regulation performance metric of the $i$th unit, which is the historical average value. $A$ is a control reward term, which equals 10 when the absolute value of the frequency deviation is smaller than 0.05.

### D. PARAMETER SELECTION
The weight coefficient in the reward function and the hyperparameters in pre-learning are designed, as shown in Table 2.
To verify the effectiveness of the proposed DCR-TD3 for MOI-AGC, a simulation test is run in the power grid of a certain province with the traditional algorithms (PI+GA, PI+PROP and PI+equal dispatch) for AGC generation power command dispatch, the regular DDPG and the TD3 as the comparison calculation examples.

The interconnected power grid system of the province has 10 units. The specific control model diagram is shown in Fig. 3, and unit parameters are shown in Table 2.

**V. SIMULATION VERIFICATION**

To verify the effectiveness of the proposed DCR-TD3 for MOI-AGC, a simulation test is run in the power grid of a certain province with the traditional algorithms (PI+GA, PI+PROP and PI+equal dispatch) for AGC generation power command dispatch, the regular DDPG and the TD3 as the comparison calculation examples.

The interconnected power grid system of the province has 10 units. The specific control model diagram is shown in Fig. 3 and unit parameters in Table 5.

**FIGURE 3.** A provincial power grid AGC control chart.

**TABLE 2.** Parameter settings.

| Parameter                     | Value     |
|-------------------------------|-----------|
| Critic learning rate          | 0.002     |
| Actor learning rate           | 0.002     |
| Discount factor               | 0.9       |
| Number of explorers           | 32        |
| Noise variance of the i-th actor | 0.05+0.005*i |
| Selection probability of experience pool 1 | 0.8       |
| Noise variance of objective action | 0.02      |
| Updating interval of strategy network | 2        |
| Size of experience 1 and 2    | 200000    |
| Weight coefficient μ₁         | 500       |
| Weight coefficient μ₂         | 0.008     |
| Weight coefficient μ₃         | 0.002     |

**FIGURE 4.** Training chart.

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

1) **PRE-LEARNING STAGE**

In the pre-learning stage, continuous sinusoidal power disturbance with a cycle of 1800s, an amplitude of 900 MW, a duration of 1800s and phase of 0.5π is applied to area A. The specific control model is shown in Fig. 3, the unit parameters are shown in Table 2, and the training chart is shown in Fig. 4.

In Fig. 4, the curves are the average reward value of every episode of each algorithm. It can be seen that the learning rate is slow and there are obvious vibrations during learning for
TABLE 3. Random step disturbance results of the provincial power grid system.

| Algorithm   | $|\Delta f|$/Hz | $|e_{act}|$/MW | $C_{act}$/% | Payment/$  |
|-------------|------------|-------------|-------------|------------|
| DCR-TD3     | 0.00635    | 26.476      | 195.937     | 10426      |
| DCR-DDPG    | 0.00641    | 26.721      | 195.785     | 10611      |
| TD3         | 0.00645    | 26.865      | 195.619     | 11107      |
| DDPG        | 0.00648    | 26.974      | 195.547     | 11259      |
| PI+GA       | 0.01554    | 60.903      | 195.064     | 28252      |
| PI+PROP     | 0.01583    | 62.003      | 194.213     | 18654      |
| PI+equal    | 0.01562    | 61.243      | 194.913     | 24947      |

the TD3 and DDPG; although the average reward values of the DCR-DDPG and the DCR-TD3 are close, the DCR-TD3 shows earlier convergence to the optimal solution and a more stable process. At the same time, as the DCR-TD3 adopts distributed training, its learning duration is significantly shorter than those of TD3 and DDPG. In addition, the average reward values of DCR-TD3 and DCR-DDPG after convergence are greater than those of the TD3 and the DDPG, meaning that distributed training can improve the quality of the solution obtained through training.
2) RANDOM STEP DISTURBANCE TEST

Random step power disturbance with an amplitude of between 1500 MW and $-400$ MW is added into the provincial power grid which consists of both coal-fired units, LNG units, oil units, hydropower units, wind turbine units as well as photovoltaic units. The test result is shown in Fig. 5 to Fig. 8 and Table 3.

It can be concluded that based on Fig. 5 to Fig. 8 and Table 3, the MOI-AGC based on DCR-TD3 delivers optimal control performance and has the smallest regulation mileage payment under random step power disturbance compared with other algorithms.

Fig. 5 to Fig. 8 show the online optimization results for MOI-AGC based on DCR-TD3, DCR-DDPG, TD3 and DDPG and traditional algorithm (PI+GA, PI+PROP and PI+equal dispatch). It can be seen from Fig. 5 the power control deviation of DCR-TD3 is much smaller than those of the PI+PROP algorithm at the same moment, and the former also has a quicker response rate, which means the actual unit regulation power output of the DCR-TD3 is closer to the actual power disturbance at the same moment. This is because the DCR-TD3 uses more quick-response regulation units (G9, G10), as shown in Fig. 6. This also results in the small change in and quick recovery of the frequency deviation of the DCR-TD3 compared with that of the traditional algorithms (Fig. 7). In addition, as the traditional algorithms adopts too many slow-response regulation units for frequency regulation, the total AGC power regulation command output by the PI controllers are greater than the power disturbance, which causes some units to have “over modulation and overshooting” and increases their regulation mileage, adding to the regulation mileage payments. By contrast, the DCR-TD3 uses mostly fast-responding units for frequency regulation and a MOI-AGC framework, which does not have the problem of instability or overshooting due to improper PI controller parameter settings. This way, not only the AGC control performance is improved, but also “overshooting” of each unit is reduced as the actual unit regulation power output is always close to the actual power disturbance; therefore the regulation mileage is reduced and so is the regulation mileage payment. As shown in Fig. 7, the regulation mileage payment of the DCR-TD3 is lower than the traditional algorithms. Based on Table 3, $|\Delta f|$, $|\varepsilon_{ACE}|$ and the regulation mileage payment of the DCR-TD3 are all the smallest; its CPS1 metric average value ($C_{\text{CPS1}}$) is the greatest.

B. SIMULATION UNDER MULTIPLE CONTINUOUS POWER DISTURBANCES

As there is a relatively great randomness and uncontrollability in wind turbine generation, photovoltaic generation as well as irregular random step power disturbance, the distributed energy output models of photovoltaic generation and wind
TABLE 4. Results of random multiple continuous power disturbances disturbance in a provincial power grid.

| Algorithm       | $|\Delta f|$/Hz | $|\Delta \text{ACE}|$/MW | $C_{\text{CPSI}}$/% | Payment/$  |
|-----------------|-------------|---------------|----------------|--------------|
| DCR-TD3         | 0.004440    | 17.868        | 196.929        | 26385        |
| DCR-DDPG        | 0.004449    | 17.908        | 196.906        | 26765        |
| TD3             | 0.004456    | 17.944        | 196.890        | 27226        |
| DDPG            | 0.004464    | 17.985        | 196.874        | 27662        |
| PI+GA           | 0.009426    | 34.862        | 196.705        | 33047        |
| PI+PROP         | 0.009754    | 35.967        | 196.145        | 27736        |
| PI+equal        | 0.009731    | 35.887        | 196.515        | 30838        |

MOI-AGC algorithms, DCR-DDPG, TD3 and DDPG, is better than that of the traditional algorithms, their frequency deviations are smaller than those of the traditional algorithms. However, as their exploration and optimum-seeking processes are not properly optimized, their control performance and economy are not as good as those of the DCR-TD3. It can be concluded that under large-scale new energy multiple continuous power disturbances, the MOI-AGC based on DCR-TD3 has a quicker response speed and better stability and economy than the AGC framework of the traditional algorithms combining PI controller and dispatch; the DCR-TD3 has better control performance and economy than the other three MOI-AGC framework algorithms and achieves a comprehensive optimum of control performance and economic benefits.

VI. CONCLUSION

In summary, the main contributions of this work are given as follows:

1) The proposed MOI-AGC based on DCR-TD3 achieves the function of “control algorithm combined with the optimization dispatch algorithm”. Through the frequency regulation market mechanism, this algorithm has the advantages of better control performance and economic benefits in the secondary regulation process of large power grid compared with traditional algorithm of “the control algorithm combined with the optimization algorithm” under the environment of multiple units and large-scale renewable energy.

2) As the algorithm is improved by using the distributed classification replay, a large number of samples can be learned. Besides, experience replay is performed by using the strategy of sampling from different experience samples with a probability, the control capability over the system is constantly improved and therefore better decisions can be made, which provides a better control effect and avoids the weaknesses of the algorithm of “the control algorithm combined with the optimization algorithm”. Compared with the DDPG and TD3 algorithms, DCR-TD3 can obtain the optimal solution more easily and has a faster convergence rate, not tending to get into a local optimum.

3) A simulation is run for a provincial power grid under two types of disturbance. The simulation result indicates that
the proposed method can significantly decrease the system frequency deviation, the total power deviation and obtain the greatest CPS1 metric value, as well as reducing the regulation mileage payment.

APPENDIX

See Table 5 and figure 11 here.

REFERENCES

[1] 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.

[2] 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, Jan. 2020, doi: 10.1109/tpwrs.2020.2966509.

[3] 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.

[4] H. Shaban, B. Vahidi, and M. Ebrahimpour, “A robust PID controller based on imperialist competitive algorithm for load frequency control of power systems,” ISA Trans., vol. 52, no. 1, pp. 88–95, Jan. 2013.

[5] 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, Aug. 2016, doi: 10.1049/iet-gtd.2016.0175.

[6] 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. 1310–1316, Jul. 2016.

[7] 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.

[8] T. Yu, Y. M. Wang, W. J. Ye, B. Zhou, and K. W. Chan, “Stochastic optimal generation command dispatch based on improved hierarchical reinforcement learning approach,” IET Gener., Transmiss. Distrib., vol. 5, no. 8, pp. 789–797, Aug. 2011.

[9] L. Xi, L. Yu, Y. Xu, S. Wang, and X. Chen, “A novel multi-agent DDQN-AD method-based distributed strategy for automatic generation control of integrated energy systems,” IEEE Trans. Sustain. Energy, early access, Dec. 9, 2019, doi: 10.1109/tste.2019.2958361.

[10] S. Mirjalili, S. M. Mirjalili, and A. Lewis, “Grey wolf optimizer,” Adv. Eng. Softw., vol. 69, no. 3, pp. 46–61, Jan. 2014.

[11] S. Mirjalili, “How effective is the Grey Wolf optimizer in training multi-layer perceptrons,” Appl. Intell., vol. 43, no. 1, pp. 150–161, 2015.

[12] T. Yu, H. Z. Wang, B. Zhou, K. W. Chan, and J. Tang, “Multi-agent correlated equilibrium Q(λ) learning for coordinated smart generation control of interconnected power grids,” IEEE Trans. Power Syst., vol. 30, no. 4, pp. 1669–1679, Jul. 2015.

[13] 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, no. 7, pp. 434–442, 2014.

[14] S. Mirjalili, “Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm,” Knowl.-Based Syst., vol. 89, pp. 228–249, Nov. 2015.

[15] S. Mirjalili and A. Lewis, “The whale optimization algorithm,” Adv. Eng. Softw., vol. 95, pp. 442–4434, May 2016.
[16] S. Mirjalili, “The ant lion optimizer,” Adv. Eng. Softw., vol. 83, no. 6, pp. 80–98, 2015.

[17] 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 2016.

[18] S. He, Q. H. Wu, and J. R. Saunders, “A group search optimizer for neural network training,” in Computational Science and Its Applications—ICCSA (Lecture Notes in Computer Science), Glasgow, U.K., vol. 3982. Berlin, Germany: Springer, 2006, pp. 934–943.

[19] X. Meng, Y. Liu, X. Gao, and H. Zhang, “A new bio-inspired algorithm: Chicken swarm optimization,” in Proc. Int. Conf. Swarm Intell. Hefei, China: Springer, 2014, pp. 86–94.

[20] S. Mirjalili, “SCA: A sine cosine algorithm for solving optimization problems,” Knowl.-Based Syst., vol. 96, pp. 120–133, Mar. 2016.

[21] 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.

[22] U.S. Federal Energy Regulatory Commission, FERC 755, Dockets RM11-7-000 AD10-11-000, Washington, DC, USA, Oct. 2011.

[23] A. D. Papalexopoulos and P. E. Andrianesis, “Performance-based pricing of frequency regulation in electricity markets,” IEEE Trans. Power Syst., vol. 29, no. 1, pp. 441–449, Jan. 2014.

[24] Y. Chen, R. Leonard, M. Keyser, and J. Gardner, “Development of performance-based two-part regulating reserve compensation on MISO energy and ancillary service market,” IEEE Trans. Power Syst., vol. 30, no. 1, pp. 142–155, Jan. 2015.

[25] 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.

[26] Z. Wang, J. Zhong, and J. Li, “Design of performance-based frequency regulation market and its implementations in real-time operation,” Int. J. Electr. Power Energy Syst., vol. 87, pp. 187–197, May 2017.

[27] T. Yu, L. Xi, B. Yang, Z. Xu, and L. Jiang, “Multiagent stochastic dynamic game for smart generation control,” J. Energy Eng., vol. 142, no. 1, Mar. 2016, Art. no. 04015012.

[28] S. Fujimoto, H. van Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods,” presented at the ICML, 2018. [Online]. Available: http://arxiv.org/abs/1802.09477

[29] T. P. Lillicrap, J. Hunt, and A. Pritzel, “Continuous control with deep reinforcement learning,” Comput. Sci., vol. 8, no. 6, p. A187, 2015.

[30] 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.

[31] 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. 112009.

[32] 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.

[33] 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, early access, Dec. 23, 2019, doi: 10.1109/TSTE.2019.2961432.

JIAWEN LI received the M.S. degree in electrical engineering from Northeast Electric Power University, Jilin, 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. His research interests include nonlinear and coordinated control theory.