energies

MDPI

Article

A Novel Reinforcement Learning Algorithm-Based Control Strategy for Grid-Configured Inverters

Xuhong Yang and Jingjian Wang *

College of Automation Engineering, Shanghai University of Electric Power, Shanghai 200090, China; yangxuhong@shiep.edu.cn

* Correspondence: 705581225@mail.shiep.edu.cn

Abstract: To address the power oscillation problem due to the introduction of inertia and damping, this paper proposes a new deep reinforcement learning algorithm based on the SD3 (Softmax Deep Double Deterministic policy gradients) algorithm for grid configuration inverter control strategy to compensate for the loss of inertia and damping in the grid. Virtual synchronous generator control, as a typical grid configuration technique, inevitably brings stability problems while providing inertia and damping support to the grid. In this paper, we first analyze the nonlinear relationship between inertia and angular velocity, and find the key parameters to maintain the power stability; then, we migrate the deep reinforcement learning strategy, and design the control strategy applicable to the virtual synchronous generator; finally, through the adaption of the key parameters, we combine the control strategy with the grid-connected inverter, and solve the problem of the excessive grid-connected power oscillation of the inverter. The effectiveness and accuracy of the control method compared with other algorithms are verified by building a simulation model in MATLAB/Simulink, which realizes the purpose of reducing power oscillation.

Keywords: deep reinforcement learning; virtual synchronous generator; power oscillations; inverter control

check for updates

Academic Editor: José Matas

Received: 17 December 2024
Revised: 21 January 2025
Accepted: 24 January 2025
Published: 27 January 2025

Citation: Yang, X.; Wang, J. A Novel Reinforcement Learning Algorithm-Based Control Strategy for Grid-Configured Inverters. *Energies* **2025**, *18*, 597. https://doi.org/10.3390/en18030597

1. Introduction

Since the twenty-first century, the global energy crisis has taken a new turn, and in order to cope with the growing shortage of energy, people have had to choose to develop new and renewable sources of energy. The development of renewable energy sources, such as wind, solar, geothermal, and tidal energy, has therefore received attention from various countries. With the development of renewable energy, the global energy structure has also changed, and distributed power generation has been widely used [1,2]. However, low inertia and poor frequency stability of the power system when renewable energy sources are connected to the grid can cause problems, and in order to solve these problems, virtual synchronous generators are proposed. By simulating the mathematical model of the synchronous generator, the virtual synchronous generator makes the inverter have the rotational inertia and damping characteristics of the synchronous generator, which reduces the frequency fluctuation while the oscillation of the system is effectively suppressed [3–5].

However, when the grid frequency or input power fluctuates greatly, the shock in the transient process may lead to damage to the device or even affect the stability of system operation. In order to suppress the fluctuation of VSG frequency and power, the solution of adjusting the rotational inertia and damping coefficient is proposed [6], in which the bang-bang algorithm of discrete variation of J (moment of inertia) and Dp (damping factor) to achieve system control is the simplest control strategy, but due to the limitation of the discrete variation, the fault tolerance and stability of the system can fail to meet the demand of the actual working conditions, and the accuracy of control is also relatively low. In [7,8], in order to solve the problems of discontinuous parameter changes and excessive power oscillations caused by the control of the bang-bang algorithm, adaptive rotational inertia and damping coefficients are proposed, and the linear relationship between them and frequency changes is established. However, the laws of the overly complex coefficient values in the strategy can have a great impact on the control effect. Refs. [9,10] proposed a nonlinear relationship between rotational inertia, damping coefficient, and angular frequency. In order to solve the problem of large fluctuation of electromagnetic power, Ref. [11] proposed an adaptive control algorithm based on RBF network (Radial Basis Function network). However, the algorithm only provides adaptive control of the rotational inertia, which makes it difficult to meet the control performance requirements in complex working conditions.

Deep reinforcement learning as a new control strategy began to be gradually applied to the field of power electronics. The deep reinforcement learning strategy observes the dynamic behavior changes of the system by continuously interacting with the system environment, and is able to adaptively optimize the VSG (Virtual Synchronous Generator) control parameters and suppress power oscillations to achieve stable operation of the power grid [12]. Ref. [13] provides a detailed analysis of deep reinforcement learning applied to grid stability control. At the same time, deep reinforcement learning based on model-free learning does not need to learn the environment model, which is relatively easier to implement and train, reduces the need for accurate mathematical models, and overcomes the problem of inaccurate environment fitting. Therefore, the model-free deep reinforcement learning-based strategy can be used as a superior control strategy to suppress the fluctuation of VSG frequency and power. In Ref. [14], a DQN (Deep Q-Network)-based control strategy is proposed for solving the unstable oscillations generated when VSGs are connected to the grid, but due to the limitations of the DQN algorithm, it can only obtain better operation results under specific operating conditions. Ref. [15] proposed a DDPG (Deep Deterministic Policy Gradient)-based VSG control strategy to achieve stable operation of the system by controlling parameter changes in real time. However, the overestimation problem may affect the stability of the algorithm, resulting in the inability to find the most suitable parameters to improve the stability of the system as much as possible. In order to solve this problem, some scholars have proposed that the TD3 (Twin Delayed Deep Deterministic policy gradient) algorithm, which is one of the model-free deep reinforcement learning algorithms, can be used [16]. However, although the TD3 algorithm improves the performance by reducing overestimation on the basis of DDPG, it will affect the performance by introducing underestimation bias. Therefore, a new algorithm, SD3, is considered to be introduced. The SD3 algorithm introduces a softmax factor based on the TD3 algorithm, which is capable of smoothing the optimization space to better improve the performance of the system.

In this paper, the SD3 algorithm is combined with VSG grid control under complex operating conditions to suppress the fluctuation of VSG frequency and power, so as to improve the stability of grid operation. The SD3 algorithm inherits the ability to solve the overestimation problem in the TD3 algorithm, which reduces the requirement for accurate mathematical models, and the algorithm outperforms the traditional reinforcement learning algorithm.

The rest of the paper will be arranged as follows. In Section 2, the basic structure of the virtual synchronous generator, the control principle, and the range of values of the rotational inertia and damping coefficient are presented. Section 3 analyzes the basic principle and implementation of the SD3 algorithm and determines the relevant dynamic variables such as action and state. Section 4 gives the results of the study to verify the effectiveness and stability of the proposed control strategy. Section 5 concludes the full paper.

2. VSG Control Principle and Parameter Determination

The VSG control scheme, shown in Figure 1, adds a deep reinforcement learning algorithm component to its conventional VSG grid-connected system, which is used to adaptively control the VSG as a way to improve the overall performance of the VSG grid-connected system and reduce power oscillations.

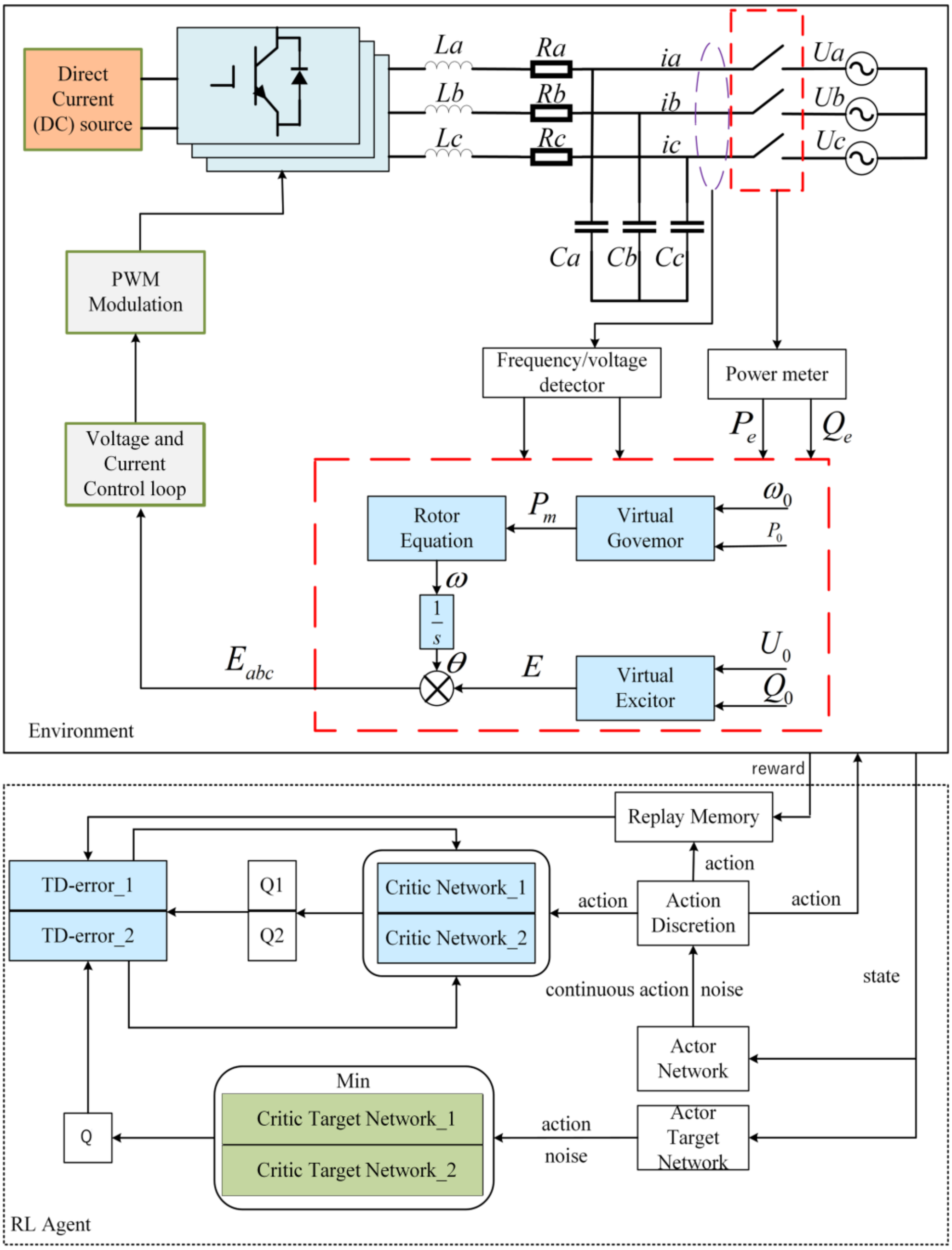


Figure 1. Overall control block diagram of improved VSG.

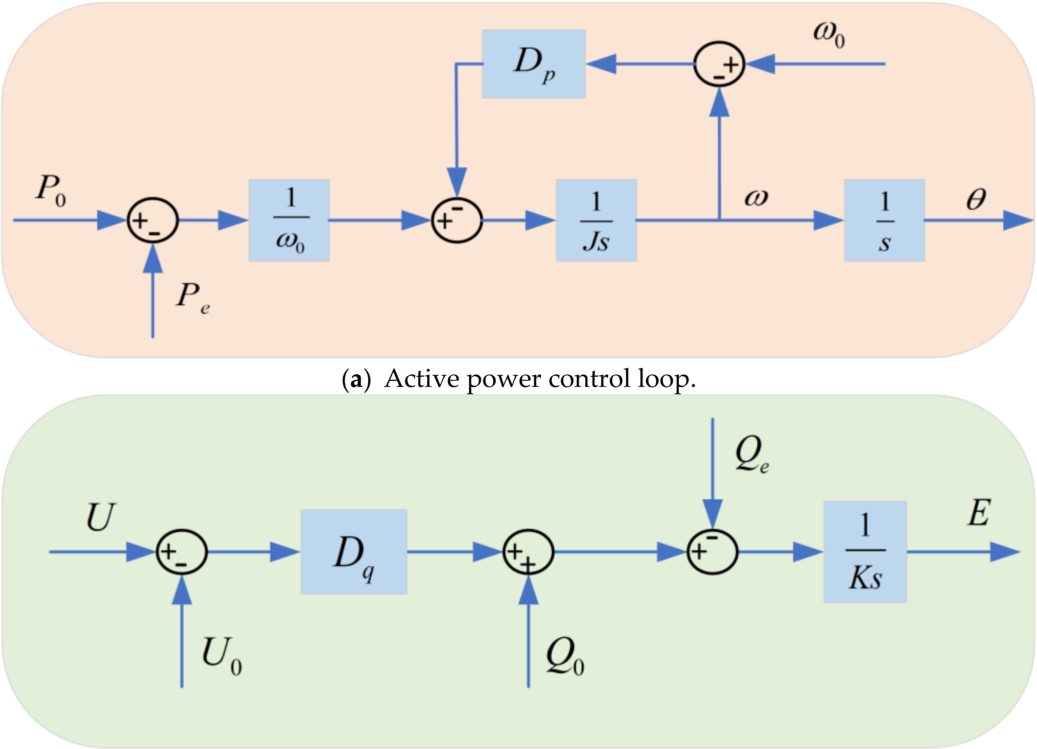
2.1. Principle of VSG Control

The VSG control algorithm primarily encompasses two main components: active and reactive control loops. The active loop is primarily comprised of the rotor equation and the virtual governor, while the reactive loop predominantly features the virtual exciter.

$$\begin{cases} J\frac{d\omega}{dt} = T_0 - T_e - D_P(\omega - \omega_0) \\ \frac{d\theta}{dt} = \omega \end{cases} \tag{1}$$

The reactive power part is illustrated in Figure 2b, which primarily emulates the excitation part of the synchronous generator, and its representation is shown in (2), where E is the internal potential output from the VSG, K is the voltage regulation coefficient, Q_0 and Q_e are reactive power value reference and output reactive power, respectively, D_q is the reactive power-voltage sag control coefficient, U and U_0 are the rated voltage and output voltage, respectively:

$$K\frac{d(E)}{dt} = Q_0 - Q_e - D_q(U - U_0) \tag{2}$$



(**a**) Active power control loop.

(**b**) Reactive power control loop.

Figure 2. VSG control loop structure.

2.2. Impact on Active Power Output Characteristics

Small-signal modelling of the VSG is carried out as in Figure 3, and the transfer function of the active loop can be obtained by approximate decoupling of the active and reactive loop parts of the figure as:

$$G_c = \frac{K_P}{J\omega_0 s^2 + D_P\omega_0 s + K_P} \tag{3}$$

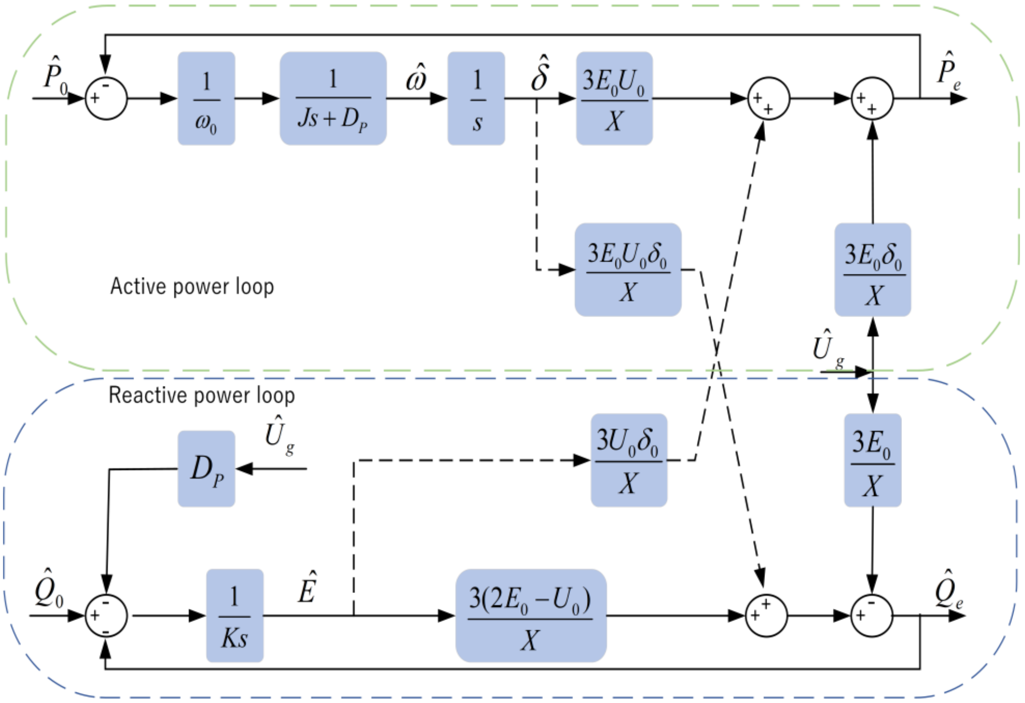


Figure 3. VSG small signal model.

In this equation, $K_p \approx 3E_0U_0/X$. Consider the intrinsic angular frequency ω_n and the damping ratio ξ of the above equation to be:

$$\begin{cases} \xi = \frac{D_P}{2}\sqrt{\frac{\omega_0}{JK_P}} \\ \omega_n = \sqrt{\frac{K_P}{J\omega_0}} \end{cases} \tag{4}$$

In general, in control engineering, with the exception of more specialized systems that do not support oscillation, control systems tend to be relatively moderately damped and at the same time can respond in a relatively short period of time. Hence, the resulting system is usually engineered as an underdamped system, with the damping ratio ξ falling within the range of (0, 1).

Then, based on the stabilization margin corresponding to the active control loop, the magnitude margin h and phase margin γ associated in the second-order link are obtained as:

$$\begin{cases} h = +\infty \\ \gamma = \arctan\left(2\xi\sqrt{\frac{1}{\sqrt{4\xi^2-1}-2\xi^2}}\right) \end{cases} \tag{5}$$

Based on (5), it can be seen that the amplitude margin tends to exceed 0. Typically for the system, the phase margin γ needs to be kept within the range of $30^\circ \sim 80^\circ$. Assuming here that $\gamma > 60^\circ$, then $\xi > 0.612$.

$$\mathrm{Re}(s_i) = -\omega_n\xi = -\frac{D_P}{2J} \leq -10 \tag{6}$$

Then, based on the cutoff frequency f_{cp}, the scope of the damping coefficient D_P can be determined. As illustrated in Figure 3, the open-loop transfer function of the active power loop can be designed as below:

$$G_{po} = \frac{3UE_o}{X\omega_o} \cdot \frac{1}{s(Js+D_P)} \tag{7}$$

$$\left|G_{po}\left(j2\pi f_{cp}\right)\right| = 1 \tag{8}$$

From (8) and (9) can be obtained the following:

$$J = \frac{D_P}{2\pi f_{cp}}\sqrt{\left(\frac{3UE_o}{2\pi f_{cp}\omega_n XD_P}\right)^2 - 1} \tag{9}$$

To ensure the validity of the equation, the expression within the square root must consistently exceed 0. The maximum cutoff frequency is usually chosen to be within 10% of the supply frequency, in which case the impact of the power loop is significantly reduced on the voltage loop. In this study, the maximum value for $f_{cp\text{max}}$ is established at 10 Hz. Consequently, the lower bound for the damping coefficient can be deduced as follows:

$$f_{cp} \leq \frac{3U_nE_n}{2\pi\omega_n XD_P} \triangleq f_{cp\text{max}} \tag{10}$$

$$D_P \geq \frac{3U_nE_n}{2\pi\omega_n Xf_{cp\text{max}}} = 11.5 \tag{11}$$

According to the EN50438 standard, the VSGs in this paper have a frequency variation equal to 1 Hz during the FM period, while the active power output from the inverter fluctuates in the interval of the rated capacity 40% $\sim$ 100%.

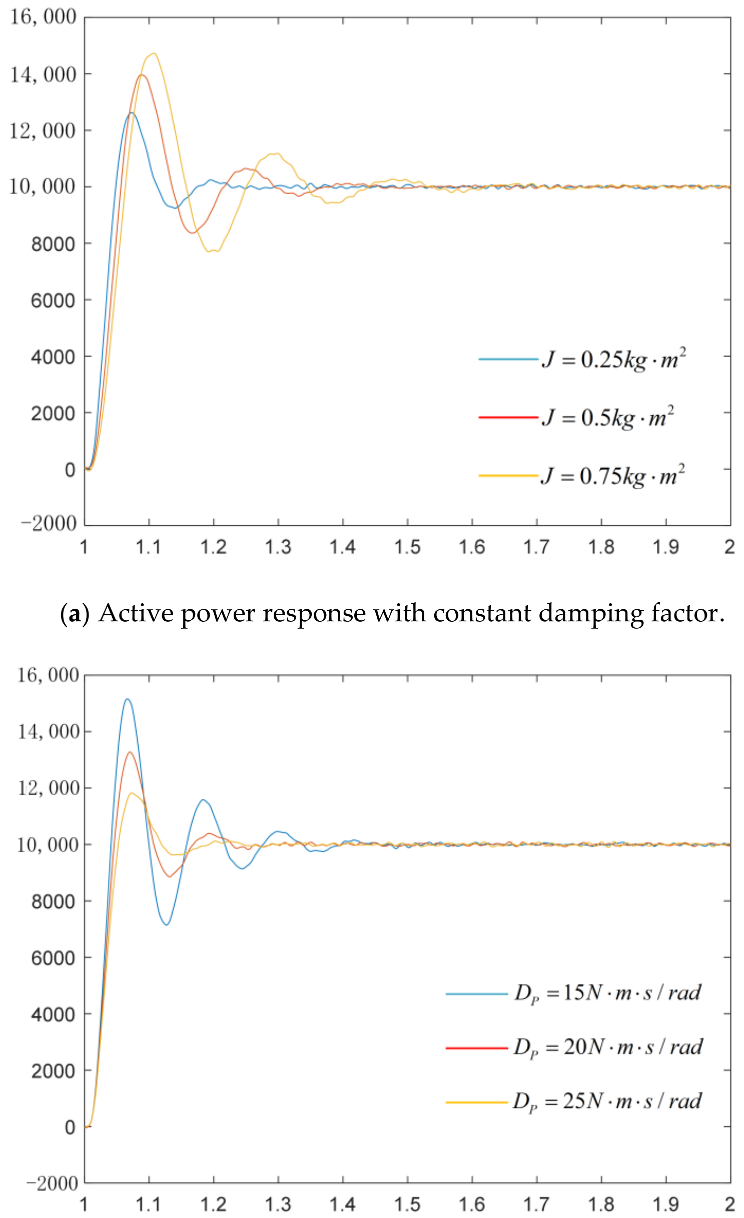
$$D_P = \frac{\triangle P}{\triangle\omega_{\text{max}}} \tag{12}$$

where $\triangle\omega_{\text{max}} = 2\pi$, then D_P obtained from (12) ranges as in (13), assuming that inverter power rating is 50 KVA.

$$10.2 \leq D_P \leq 25.4 \tag{13}$$

By counting, it can be obtained that the rotational inertia J is in the range of [0.25, 2.9] and the damping coefficient D_P is in the range of [10.1, 25.3].

For a given active power of 10 KVA and reactive power of 5 KVA, the active power dynamic performance of the system is completely determined by the moment of inertia J and damping coefficient D_P. The active power dynamic response trajectory of the VSG can be plotted for a known range of different moment of inertia and damping coefficients, as shown in Figure 4. The analysis of Figure 4a shows that, assuming that the damping coefficient D_P is kept constant, the rotational inertia J is inversely proportional to the damping ratio ξ and positively proportional to the overshooting amount σ. The larger J is, the smaller ξ is, the larger σ is, and the longer the regulation time t_s is; from Figure 4b, it can be deduced that, assuming that J is kept constant, a positive proportionality relationship between D_P and ξ also exists and an inverse relationship with σ. The larger D_P is, the larger ξ is, the smaller the overshooting amount σ is and the regulation time t_s is also smaller. It can be concluded that the moment of inertia determines the oscillation frequency during the dynamic response of the VSG active power, and the damping coefficient determines the decay rate of the dynamic response of the VSG active power; therefore, the adaptive control of the moment of inertia and damping coefficient can make system operation become more stable, and also make the system more robust.



(**a**) Active power response with constant damping factor.

(**b**) Active power response with constant moment of inertia.

Figure 4. VSG active power dynamic response.

2.3. Impact on Corner Frequency Output Characteristics

In order to analyze the effect of damping coefficient and rotational inertia on the output characteristics of the system, the amount of overshooting of the angular frequency is observed for a given active power of 15 KW and reactive power of 10 KWA.

From Figure 5, it can be seen that increasing the rotational inertia and damping coefficient at the same time can reduce the amount of overshooting of the angular frequency, and the smaller the rotational inertia, the smaller the regulation time; however, because the damping coefficient is too small the system will produce a larger overshooting, and small damping can not be quickly adjusted to the overshooting fluctuations; therefore, an appropriate increase in the damping can reduce the regulation time, but with the increase in the damping coefficient, the regulation time will be increased due to the system's response speed to the slow-downs and increases. Therefore, the larger the moment of inertia is, the smaller the angular frequency fluctuation is, the more stable the system is, but if it is too large, it will lead to poor system performance. Under the premise that $T_0 - T_e - J\frac{d\omega}{dt}$ is constant, the larger the damping coefficient is, the smaller the overshoot of angular frequency is, but too large a damping will lead to slower response of the system.

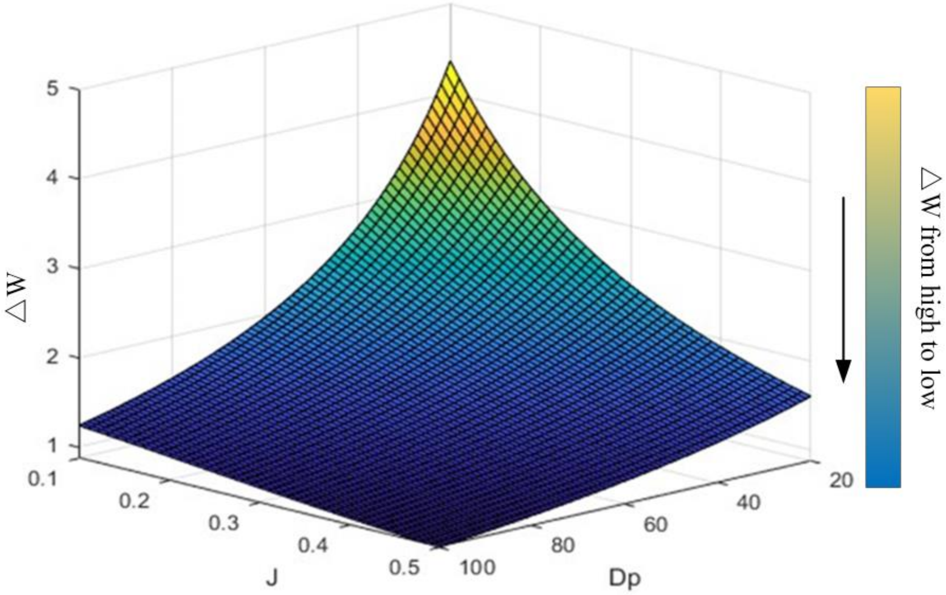


Figure 5. Variation of angular frequency for different moments of inertia and damping coefficients.

3. Reinforcement Learning

With no precise model and uncertain environments, model-free techniques have proven to be very efficient and beneficial. Reinforcement learning (RL) is a clever strategy that is unaffected by model parameters or the surrounding environment. The learning process in RL is comparable to human learning tendencies [16,17]. Through activities, the individual (agent) interacts with the environment, gaining experience and becoming an expert on the individual task over time. Actions that can enhance outcomes are emphasized in reinforcement learning. Every action taken by the agent creates a new state. The agent's reward for a particular action is determined by the likelihood of the following state being favorable. Figure 6 depicts the RL building blocks.

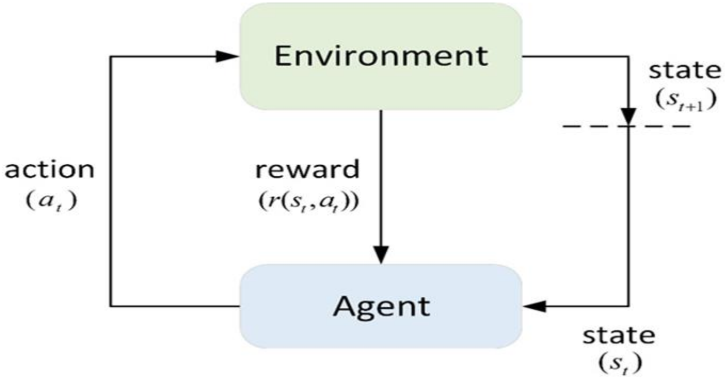


Figure 6. Block diagram of reinforcement learning.

3.1. SD3 Algorithm Logical Framework

The SD3 algorithm chosen in this paper is an improvement of the Deep Deterministic Policy Gradient algorithm, and a diagram of the basic principle of the algorithm is as Figure 7.

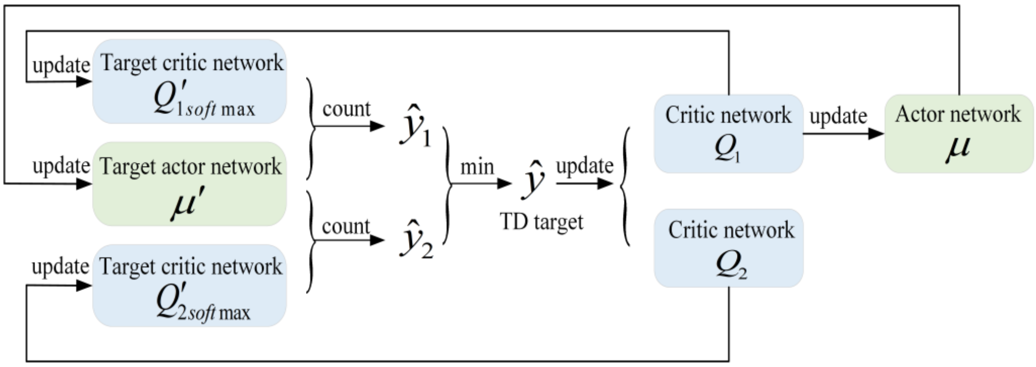


Figure 7. Basic flow of the SD3 algorithm.

The SD3 algorithm follows the truncated double Q-value learning strategy of the TD3 algorithm, and updates the cubic network in a similar way to double Q-learning by learning two Q-value functions [18]. The max operation in deep Q-network algorithms leads to the problem of overestimation of Q-values. This problem also exists in deep deterministic policy gradient algorithms, because $Q(s,a)$ in deep deterministic policy gradient algorithms is updated in the same way as in deep Q-network algorithms, whereas there is an error in the estimation of Q-values when using a tool such as neural network as a function approximator to deal with a complex problem:

$$Q^{appox}(s',\hat{a}) = Q^{target}(s',\hat{a}) + Y_{s'}^{\hat{a}} \tag{14}$$

where $Y_{s'}^{\hat{a}}$ is zero-mean noise, but the use of the max operation leads to an error between Q^{appox} and Q^{target}. Denoting the error as Z_s, we can derive:

$$Z_s \overset{def}{=\!=} \gamma(\max_{\hat{a}} Q^{approx}(s',\hat{a}) - \max_{\hat{a}} Q^{target}(s',\hat{a})) \tag{15}$$

Consider that in the noise term, some Q-values may be small while others may be large. The max operation always chooses the largest Q-value for each state, which causes the algorithm to be unusually sensitive to the corresponding Q-value of the overestimated action. In this case, it will result in creating an overestimation problem.

The TD3 algorithm introduces double Q-learning into the deep deterministic policy gradient algorithm by building a network of two Q-values to estimate the value of the next state:

$$Q_{\theta'_1}(s',a') = Q_{\theta'_1}(s',\pi_{\phi_1}(s')) \tag{16}$$

$$Q_{\theta'_2}(s',a') = Q_{\theta'_2}(s',\pi_{\phi_1}(s')) \tag{17}$$

In the TD3 algorithm, the Bellman equation is computed using the smaller of the two Q values:

$$Y_1 = r + \gamma \min_{i=1,2} Q_{\theta'_i}(s',\pi_{\phi_i}(s')) \tag{18}$$

Using truncated dual Q-learning, the valuation of the target network does not introduce excessive estimation error to the Q-learning objective. However, this updating pattern may lead to underestimation, so Ref. [19] investigated the relationship between underestimation and the target value in the TD3 algorithm, and in order to solve the underestimation problem, a softmax factor is considered to be introduced in the target value:

$$\begin{aligned} Y_1 &= r + \gamma soft\max_{\beta} Q'_{\theta} \\ Q'_{\theta} &= \min_{i=1,2} Q_{\theta'_i}(s',\pi_{\phi_i}(s')) \end{aligned} \tag{19}$$

Delayed policy update: During the update process, the policy network is updated less frequently than the Q-value network. The SD3 algorithm reduces the frequency of updating the policy network, which is updated only after the value network has been updated d times. This type of policy updating allows the estimation of the Q-value function to have a smaller variance, resulting in higher-quality policy updates [20].

Target strategy smoothing: Noise is added to the output action of the target strategy as a way of smoothing the estimation of the Q function to avoid overfitting [21].

One problem with deterministic strategies is the potential overfitting of this class of methods for narrow peak estimates in the value space [20]. In the SD3 algorithm, similar actions should have similar value estimates, so it is reasonable to fuzzy fit the values of a small region around the target action:

$$y = r + E_{\varepsilon}[Q_{\theta'}(s', \pi_{\phi'}(s') + \varepsilon)] \tag{20}$$

Overfitting is avoided by adding truncated normally distributed noise as regularization in each action to smooth the calculation of Q-values. The corrected update is as follows:

$$y = r + \gamma soft\mathrm{max}_{\beta} Q_{\theta'}(s', \pi_{\phi'}(s') + \varepsilon), \varepsilon \sim clip(N(0, \sigma), -c, c) \tag{21}$$

Algorithm 1 is shown as follows:

Algorithm 1: SD3

```
Initialize critic networks $Q_{\theta_1}$, $Q_{\theta_2}$, and actor network $\pi_\Phi$ with random parameters $\theta_1$, $\theta_2$, $\Phi$
Initialize target networks $\theta'_1 \leftarrow \theta_1$ , $\theta'_2 \leftarrow \theta_2$ , $\Phi' \leftarrow \Phi$
Initialize replay buffer $D$
for $t = 1$ to $T$ do
    Select action with exploration noise $A_t \sim \pi_\Phi(S_t) + \epsilon, \epsilon \sim N(0, \sigma)$
    Accept reward $R_t$ and new state $S_{t=1}$
    Store transition tuple $(S_t, A_t, R_t, D_t, S_{t+1})$ in $D$
for $i$ = 1, 2 do
    Sample mini-batch of $N$ transitions $(S_t, A_t, R_t, D_t, S_{t+1})$ from $D$
    $\widetilde{a}_{t+1} \leftarrow \pi_{\Phi'}(S_{t+1}) + \epsilon, \epsilon \sim clip\left(N\left(0, \widetilde{\sigma}, -c, c\right)\right)$
    $\hat{Q}\left(s', \hat{a'}\right) \leftarrow min_{i=1,2} Q_{\theta'_i}\left(S_{t+1}, \widetilde{a}_{t+1}\right)$
    $softmax_\beta Q'_\theta \leftarrow E_{\hat{a'} \sim p}\left[\frac{\exp\left(\beta\hat{Q}(s',a')\hat{Q}(s',a')\right)}{p\left(\hat{a'}\right)}\right] / E_{\hat{a'} \sim p}\left[\frac{\exp\left(\beta\hat{Q}(s',a')\right)}{p\left(\hat{a'}\right)}\right]$
    $y \leftarrow r + \gamma(1-d) softmax_\beta Q'_\theta$
    Update Critic network $\theta_i \leftarrow argmin_{\theta_i} N^{-1} \sum \left(y - Q_{\theta_i}(S_t, A_t)\right)^2$
    Update $\Phi$ by the deterministic policy gradient:
    $\nabla_\Phi J(\Phi) = N^{-1} \sum \nabla_a Q_{\theta_i}(S_t, A_t)|_{A_t = \pi_\Phi(S_t)} \nabla_\Phi \pi_\Phi(S_t)$
    Update  target networks:
    $\hat{\theta}_i \leftarrow \rho\theta_i + (1-\rho)\hat{\theta}_i$
    $\hat{\Phi} \leftarrow \rho\Phi + (1-\rho)\hat{\Phi}$
end if
end if
```

3.2. Algorithmic Tasking

In a reinforcement learning task, there are three key elements that need to be defined for it, namely: observation states, actions, and rewards. The observation set selected in this paper is as follows:

$$S_t \in \left\{ \Delta\omega, \Delta P, \Delta Q, \frac{d\omega}{dt} \right\} \tag{22}$$

where $\Delta\omega = \omega_0 - \omega$, $\Delta P = P_0 - P_e$, and $\Delta Q = Q_0 - Q_e$, denote the difference between the angular frequency, active power, and reactive power, respectively, and their reference values.

In this paper, the rotational inertia J and the damping coefficient D_p are chosen as its actions. The set of actions is thus set as:

$$A_t \in \left\{ J, D_p \right\} \tag{23}$$

where $J \in [J_{\min}, J_{\max}]$,$D_p \in \left[D_{p\min}, D_{p\max}\right]$. The initial values for the actions are set to be J_0 and D_{p0}.

$$r_t = -l_\omega C(\omega_t) - l_P C(P_t) - l_Q(Q_t) \tag{24}$$

where T is the total time and l_ω, l_P, and l_Q are weighting factors greater than 0. The values of the reward function are all negative in order to better ensure the stability and security of the system.

The specific design for (24) is as follows:

$$C(\omega_t) = \begin{cases} \varphi_\omega \alpha_w & \alpha_w \le \alpha_\omega^{\max} \\ \rho_w & \alpha_\omega > \alpha_\omega^{\max} \end{cases} \tag{25}$$

$$\alpha_w = |\omega_0 - \omega| \tag{26}$$

where φ_ω is a small penalty coefficient and ρ_w is a large penalty coefficient. Observing (25) and (26), a small penalty term is designed at $\alpha_w \le \alpha_\omega^{\max}$ to make the difference between ω_0 and ω as small as possible, while when $\alpha_\omega > \alpha_\omega^{\max}$, it can be assumed that the results of this round of training do not meet expectations, so a large penalty term is designed and the round is terminated.

$$C(P_t) = \varphi_P |P_0 - P_e| \tag{27}$$

$$C(Q_t) = \varphi_Q |Q_0 - Q_e| \tag{28}$$

For (27) and (28), φ_P and φ_Q are the penalty coefficients for active and reactive power, respectively. Since the dynamic performance does not depend on one moment of penalty alone, and the accumulation of penalties over a long period of time can give the system better stability, a discount factor β is introduced, and the final reward function of the system is as follows:

$$R = \sum_{n=t}^{T} \beta^{n-t} r_n \tag{29}$$

where T is the total time. For the design of each coefficient of the reward function, see Table 1.

Table 1. Reward function coefficients.

Symbol	Value
l_p	0.25
l_ω	0.25
l_q	0.25
$\alpha_\omega^{\max}$	$2\pi \times 0.6$
φ_ω	10
ρ_ω	300
β	0.9

4. Results

The paper adopts the actor-critic structure of the SD3 algorithm, as illustrated in Figure 8. The critic network encompasses state pathways, action pathways, and a shared pathway. In the case of the actor network, it receives observations as inputs and produces corresponding actions as outputs. Notably, within this framework, “ReLU” and “Tanh” serve as standard activation functions for neurons, well-established in the design of deep neural networks. The scaling layer plays a pivotal role in vector scaling, while the fully connected layer performs matrix multiplication on the input, subsequently adding a bias vector. A comprehensive breakdown of the SD3 algorithm’s parameters can be found in Table 2.

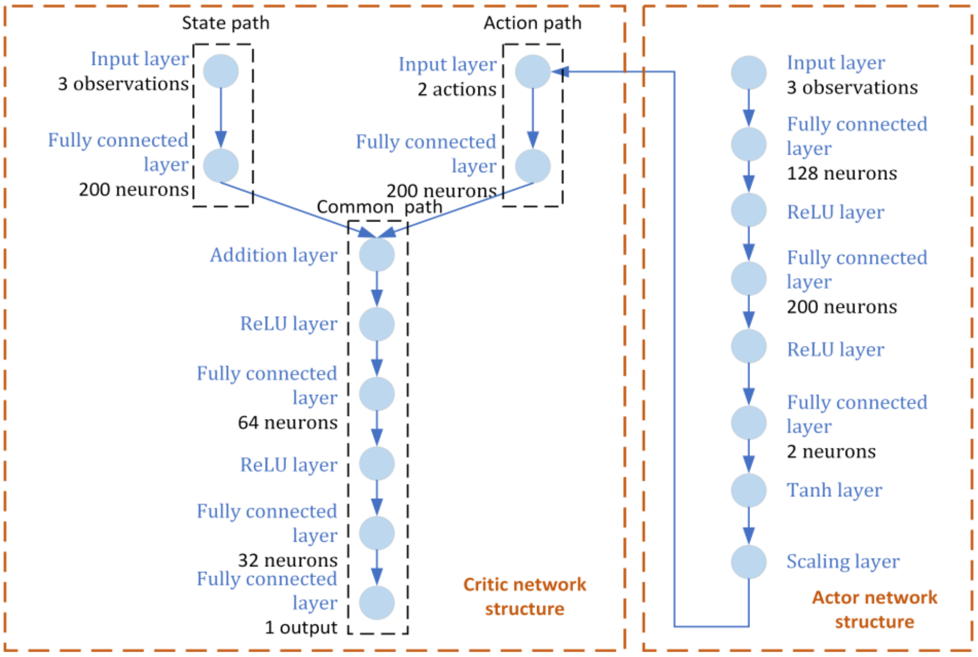


Figure 8. Actor–critic structure.

As an efficient power generation technology, the application of fuel cells in the field of grid-connection power generation has gradually increased in recent years. The cell voltage V_{cell} of a fuel cell can be given by the following equation:

$$V_{cell} = E_n - V_{act} - V_{ohm} - V_{con} \tag{30}$$

where E_n is the thermodynamic electric potential; V_{act} is the activation polarization overpotential; V_{ohm} is the ohmic polarization overpotential; and V_{con} is the concentration polarization overpotential. Since the output voltage of a single fuel cell is relatively low, this paper uses multiple cells to form a fuel cell stack to increase its output voltage.

Table 2. RL training parameter.

Symbol	Value
Simulation time	4
Sample time	0.005
Discount factor	0.995
Mini-Batch size	256
Reply buffer size	100,000
Learning rate for actor	0.0001
Learning rate for critic	0.0001
Training episodes	500
Target smooth factor	0.005
Target update frequency	10

In this section, a VSG-PEM grid-connection system is proposed based on fuel cells, in which the SD3 algorithm is compared with RBF neural network and traditional adaptive control for grid-connected performance, and the VSG model is developed in this paper in the MATLAB/Simulink 2023a platform for simulation. In this case, the parameters of the fuel cell are shown in Table 3.

Table 3. Fuel cell simulation model parameter settings.

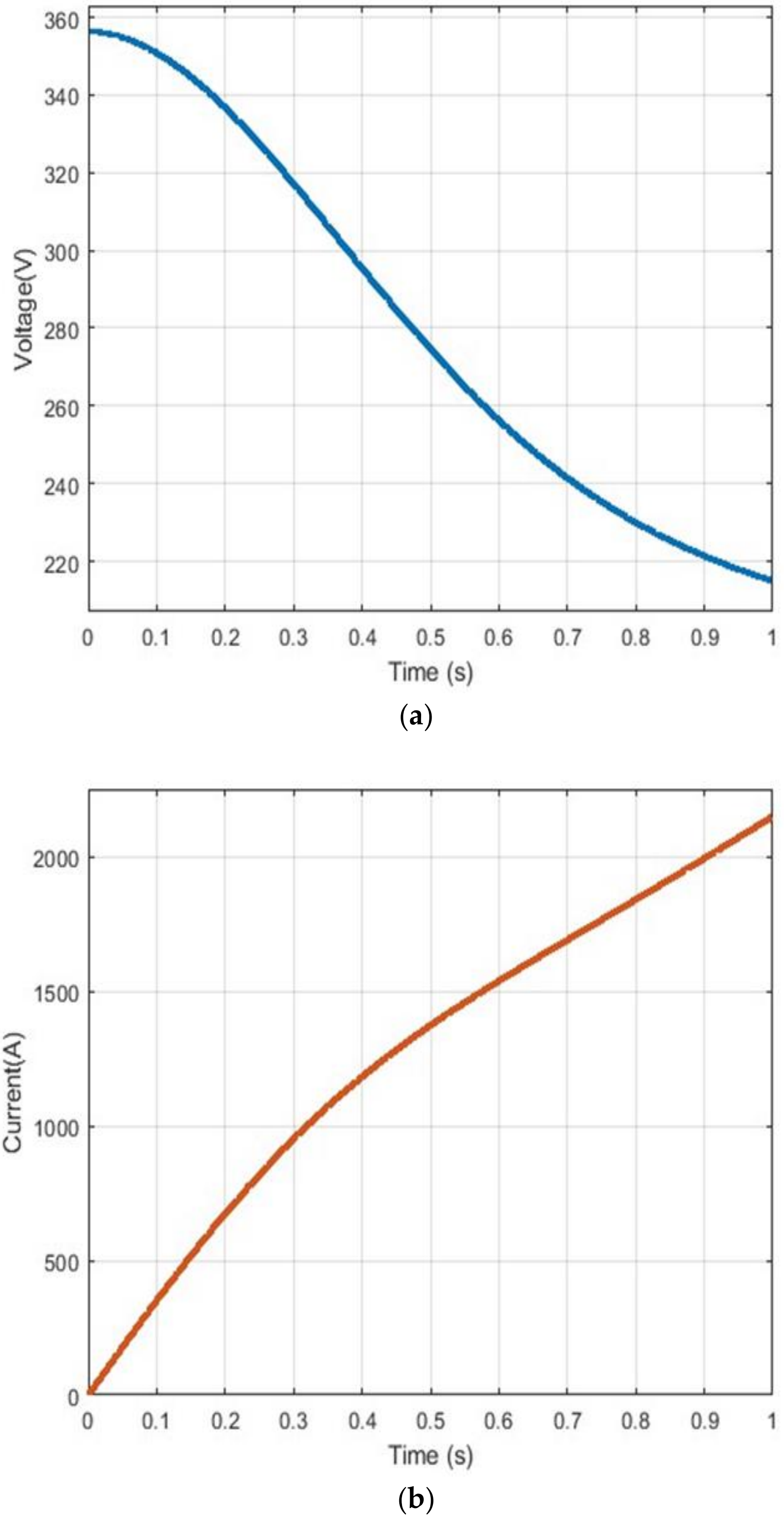
Parameters	Symbol	Value
Number of batteries	-	700
Proton exchange membrane thickness	µm	0.0178
Anode hydrogen partial pressure	MPa	1
Cathode oxygen partial pressure	MPa	0.2
Reference temperature	K	333
Effective activation area of single cell	cm^2	64

Figure 9 shows the voltage and current curves of the fuel cell, from which it can be analyzed that the volt-ampere characteristics of the fuel cell stack show a negative correlation, i.e., as the output current of the stack increases, the output voltage of the stack gradually decreases.

In case of supply–demand imbalance, if the active power command is increased from 100 kW to 150 kW within 0.6 s of the initial run, then the reactive power command is increased from 5 kvar to 10 kvar within 1.1 s. Figure 10 shows the training results of rotational moment of inertia and damping coefficients when the SD3 algorithm is used.

Figure 11 shows the variation of active power when the initial operating active load is increased from 100 kW to 150 kW in 0.6 s. In this case, the overshoot of active power is 5.3% for the SD3 algorithm and 7.3% and 8% for the TD3 algorithm and the DDPG algorithm, respectively. It can also be seen from the figure that compared with the TD3 algorithm and the DDPG algorithm, the SD3 algorithm has faster response speed, shorter adjustment time, and the stability of the system has been improved more.

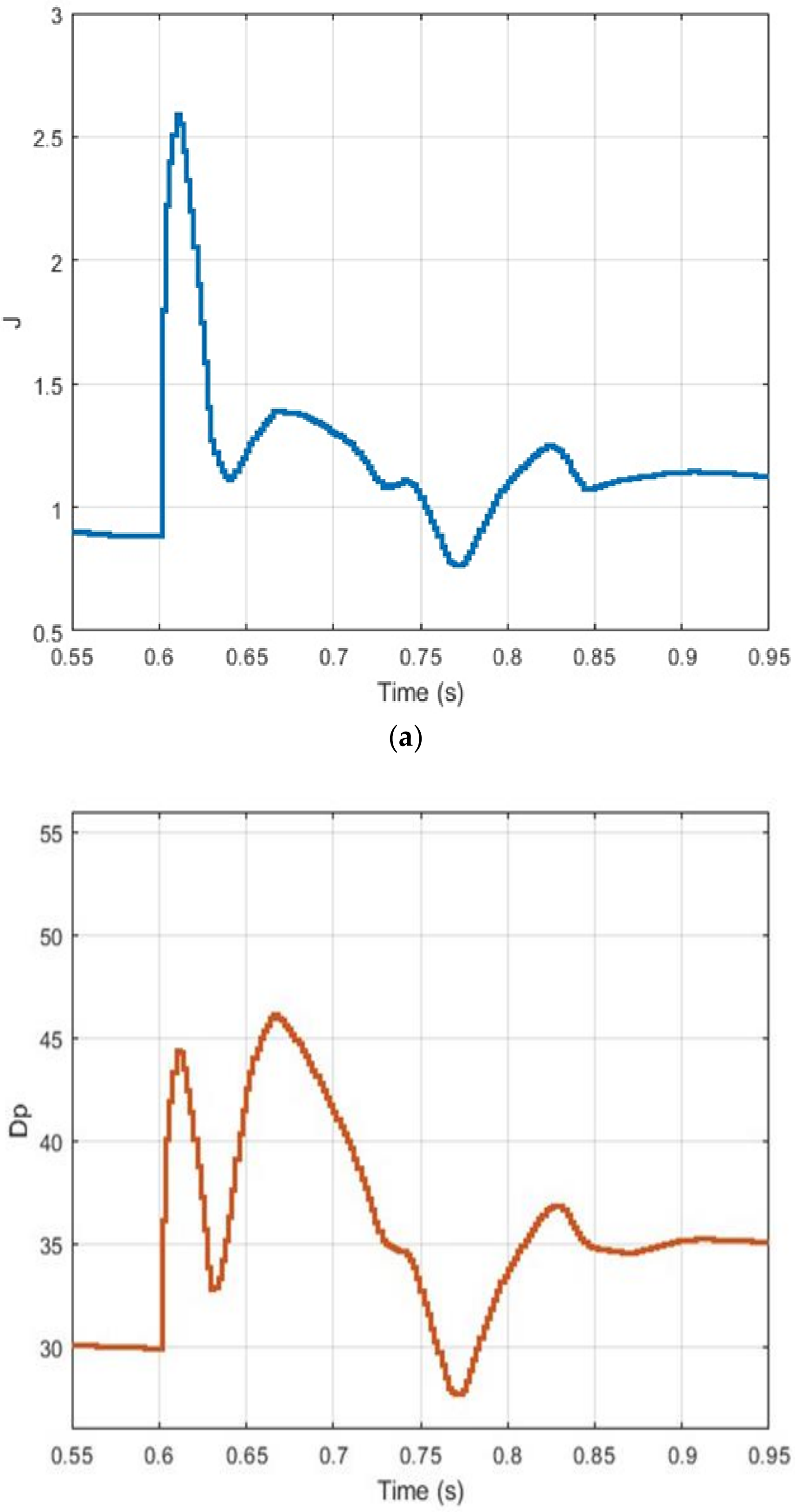
As can be seen from Figure 12, compared with the other two control strategies, the SD3 algorithm performs better than the TD3 algorithm and the DDPG algorithm in suppressing the angular frequency fluctuation caused by the sudden change in active power or the angular frequency fluctuation caused by the sudden change in reactive power. Among them, when the reactive power mutation occurs, the fluctuation of the corner frequency under the SD3 algorithm is smaller, and the recovery time of the corner frequency is about 0.05 s, which is also better than the TD3 algorithm and the DDPG algorithm.



(**a**)

(**b**)

Figure 9. Voltage and current graphs of fuel cells. (**a**) Voltage–time. (**b**) Current–time.



(**a**)

(**b**)

Figure 10. RL results of inertia and damping coefficient. (**a**) Moment of inertia. (**b**)Damping coefficient.

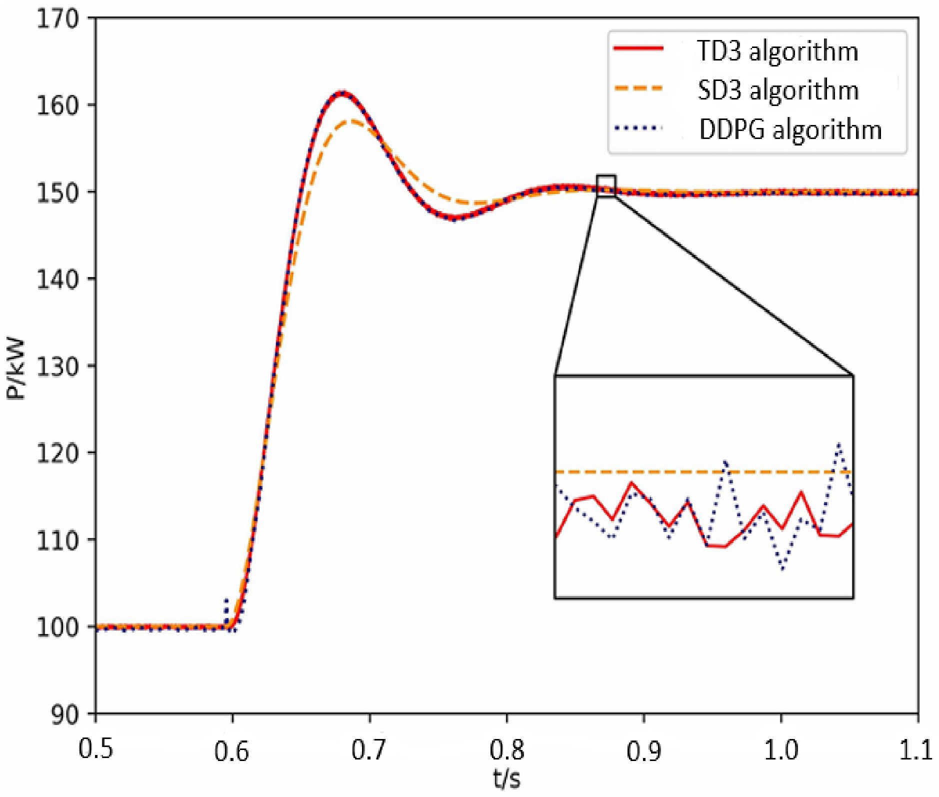


Figure 11. Performance comparison of the SD3 algorithm with other algorithms for active changes.

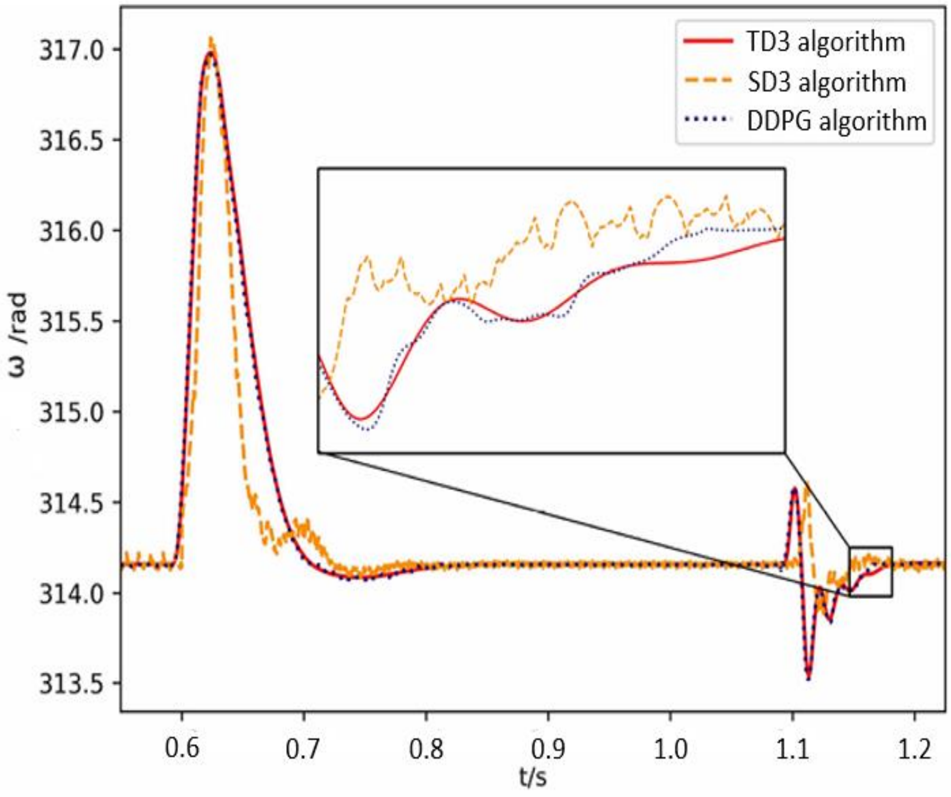


Figure 12. Comparison of the performance of the SD3 algorithm in suppressing angular frequency fluctuation.

5. Conclusions

In order to overcome the problems of inertia and damping loss in the power grid, this paper proposes a control strategy based on the SD3 algorithm, which is firstly applied to solve the problem of compensating for the rotational inertia and damping coefficients of VSGs under complex working conditions. The control strategy can effectively suppress the frequency and power fluctuations. The main work and conclusions of this paper are as follows:

(1) According to the four aspects of VSG stability margin, frequency regulation time, power circuit cutoff frequency, and grid-connected standard, the reasonable value ranges of virtual inertia and damping coefficient are determined to ensure the stability of the system.

(2) The control mechanism of the SD3 algorithm in deep reinforcement learning is deeply analyzed, and an artificial intelligence VSG control strategy based on the SD3 algorithm is proposed. Simulation results show that the strategy overcomes the shortcomings of the traditional VSG. It can suppress the frequency overshoot and improve the response speed.

(3) The excellent performance of the SD3 algorithm in dealing with continuous control tasks is verified, but its problems such as more complicated computation and more difficult adjustment of hyperparameters need to be solved by subsequent continuous research.

Author Contributions: X.Y. established the original conception, provided technical guidance and checked the data. J.W. modeled the system, designed the algorithms of control strategy and wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Science Foundation of China (51777120) and Shanghai 2021 "Science and Technology Innovation Action Plan" Science and Technology Support for Carbon Neutralization (the first batch) (21DZ1207502).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Wang, T.; O'Neill, D.; Kamath, H. Dynamic Control and Optimization of Distributed Energy Resources in a Microgrid. *IEEE Trans. Smart Grid* **2015**, *6*, 2884–2894. [CrossRef]
2. Ammar, M. A Flicker Allocation Scheme for MV Networks With High Penetration of Distributed Generation. *IEEE Trans. Power Deliv.* **2016**, *31*, 400–401. [CrossRef]
3. Qu, Z.; Peng, J.C.-H.; Yang, H.; Srinivasan, D. Modeling and Analysis of Inner Controls Effects on Damping and Synchronizing Torque Components in VSG-Controlled Converter. *IEEE Trans. Energy Convers.* **2021**, *36*, 488–499. [CrossRef]
4. An optimal coordination control strategy of micro-grid inverter and energy storage based on variable virtual inertia and damping of vsg. *Chin. J. Electr. Eng.* **2017**, *3*, 25–33. [CrossRef]
5. Zhou, L.; Liu, S.; Chen, Y.; Yi, W.; Wang, S.; Zhou, X.; Wu, W.; Zhou, J.; Xiao, C.; Liu, A. Harmonic Current and Inrush Fault Current Coordinated Suppression Method for VSG Under Non-ideal Grid Condition. *IEEE Trans. Power Electron.* **2021**, *36*, 1030–1042. [CrossRef]
6. Li, D.; Zhu, Q.; Lin, S.; Bian, X.Y. A Self-Adaptive Inertia and Damping Combination Control of VSG to Support Frequency Stability. *IEEE Trans. Energy Convers.* **2017**, *32*, 397–398. [CrossRef]
7. Alipoor, J.; Miura, Y.; Ise, T. Power System Stabilization Using Virtual Synchronous Generator With Alternating Moment of Inertia. *IEEE J. Emerg. Sel. Top. Power Electron.* **2015**, *3*, 451–458. [CrossRef]

8. Li, J.; Wen, B.; Wang, H. Adaptive Virtual Inertia Control Strategy of VSG for Micro-Grid Based on Improved Bang-Bang Control Strategy. *IEEE Access* **2019**, *7*, 39509–39514. [CrossRef]
9. Mentesidi, K.; Garde, R.; Aguado, M.; Rikos, E. Implementation of a fuzzy logic controller for virtual inertia emulation. In Proceedings of the 2015 International Symposium on Smart Electric Distribution Systems and Technologies (EDST), Vienna, Austria, 8–11 September 2015; pp. 606–611. [CrossRef]
10. Wang, Z.; Meng, F.; Zhang, Y.; Wang, W.; Li, G.; Ge, J. Cooperative Adaptive Control of Multi-Parameter Based on Dual-Parallel Virtual Synchronous Generators System. *IEEE Trans. Energy Convers.* **2023**, *38*, 2396–2408. [CrossRef]
11. Yao, F.; Zhao, J.; Li, X.; Mao, L.; Qu, K. RBF Neural Network Based Virtual Synchronous Generator Control With Improved Frequency Stability. *IEEE Trans. Ind. Inform.* **2021**, *17*, 4014–4024. [CrossRef]
12. Wang, Z.; Yu, Y.; Gao, W.; Davari, M.; Deng, C. Adaptive, Optimal, Virtual Synchronous Generator Control of Three-Phase Grid-Connected Inverters Under Different Grid Conditions—An Adaptive Dynamic Programming Approach. *IEEE Trans. Ind. Informatics* **2022**, *18*, 7388–7399. [CrossRef]
13. Yang, Q.; Yan, L.; Chen, X.; Chen, Y.; Wen, J. A Distributed Dynamic Inertia-Droop Control Strategy Based on Multi-Agent Deep Reinforcement Learning for Multiple Paralleled VSGs. *IEEE Trans. Power Syst.* **2023**, *38*, 5598–5612. [CrossRef]
14. Wu, W.; Guo, F.; Ni, Q.; Liu, X.; Qiu, L.; Fang, Y. Deep Q-Network based Adaptive Robustness Parameters for Virtual Synchronous Generator. In Proceedings of the 2022 IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific), Haining, China, 28–31 October 2022; pp. 1–4. [CrossRef]
15. Chen, L.; Tang, J.; Qiao, X.; Chen, H.; Zhu, J.; Jiang, Y.; Zhao, Z.; Hu, R.; Deng, X. Investigation on Transient Stability Enhancement of Multi-VSG System Incorporating Resistive SFCLs Based on Deep Reinforcement Learning. *IEEE Trans. Ind. Appl.* **2024**, *60*, 1780–1793. [CrossRef]
16. Kaifang, W.; Bo, L.; Xiaoguang, G.; Zijian, H.; Zhipeng, Y. A learning-based flexible autonomous motion control method for UAV in dynamic unknown environments. *J. Syst. Eng. Electron.* **2021**, *32*, 1490–1508. [CrossRef]
17. Li, Y.; Hu, X.; Zhuang, Y.; Gao, Z.; Zhang, P.; El-Sheimy, N. Deep Reinforcement Learning (DRL): Another Perspective for Unsupervised Wireless Localization. *IEEE Internet Things J.* **2020**, *7*, 6279–6287. [CrossRef]
18. Luo, X.; Wang, Q.; Gong, H.; Tang, C. UAV Path Planning Based on the Average TD3 Algorithm With Prioritized Experience Replay. *IEEE Access* **2024**, *12*, 38017–38029. [CrossRef]
19. Pan, L.; Cai, Q.; Huang, L. Softmax Deep Double Deterministic Policy Gradients. *arXiv* **2020**, arXiv:2010.09177. [CrossRef]
20. Wang, H.; Feng, L.; Zhang, Y.; Zhou, J.; Du, H. Human-machine Authority Allocation in Indirect Cooperative Shared Steering Control with TD3 Reinforcement Learning. *IEEE Trans. Veh. Technol.* **2024**, *73*, 7576–7588. [CrossRef]
21. Khalid, J.; Ramli, M.A.; Khan, M.S.; Hidayat, T. Efficient Load Frequency Control of Renewable Integrated Power System: A Twin Delayed DDPG-Based Deep Reinforcement Learning Approach. *IEEE Access* **2022**, *10*, 51561–51574. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.