energies

MDPI

Article

Twin-Delayed Deep Deterministic Policy Gradient for Low-Frequency Oscillation Damping Control

Qiushi Cui [1,2], Gyoungjae Kim [1] and Yang Weng [1,*]

[1] School of Electrical, Computer and Energy Engineering, Arizona State University, 551 East Tyler Mall, Tempe, AZ 85281, USA; qcui5@asu.edu (Q.C.); gkim50@asu.edu (G.K.)
[2] School of Electrical Engineering, Chongqing University, Chongqing 400044, China
* Correspondence: yweng2@asu.edu; Tel.: +1-480-965-8202

Abstract: Due to the large scale of power systems, latency uncertainty in communications can cause severe problems in wide-area measurement systems. To resolve this issue, a significant amount of past work focuses on using emerging technology, including machine learning methods such as Q-learning, for addressing latency issues in modern controls. Although the method can deal with the stochastic characteristics of communication latency, the Q-values can be overestimated in Q-learning methods, leading to high bias. To address the overestimation bias issue, we redesign the learning structure of the deep deterministic policy gradient (DDPG). Then we develop a damping control twin-delayed deep deterministic policy gradient method to handle the damping control issue under unknown latency in the power network. The purpose is to address the damping control issue under unknown latency in the power network. This paper will create a novel reward algorithm, taking into account the machine speed deviation, the episode termination prevention, and the feedback from action space. In this way, the system optimally damps down frequency oscillations while maintaining the system's stability and reliable operation within defined limits. The simulation results verify the proposed algorithm in various perspectives, including the latency sensitivity analysis under high renewable energy penetration and the comparison with conventional and machine learning control algorithms. The proposed method shows a fast learning curve and good control performance under varying communication latency.

Keywords: latency; twin-delayed deep deterministic policy gradient; damping control; wide-area measurement systems; low-frequency oscillations

Citation: Cui, Q.; Kim, G.; Weng, Y. Twin-Delayed Deep Deterministic Policy Gradient for Low-Frequency Oscillation Damping Control. *Energies* **2021**, *14*, 6695. https://doi.org/10.3390/en14206695

Academic Editors: Giuliano Armano, Paolo Attilio Pegoraro and Elyas Rakhshani

Received: 16 July 2021
Accepted: 27 September 2021
Published: 15 October 2021

1. Introduction

Inter-area low-frequency oscillations cause significant challenges to reliable control and economic operations in a typical cyber-physical system such as transmission networks. For example, if the inter-area oscillation has poor damping, it will cause catastrophic disturbances, such as forming multiple outages, which lead to widespread oscillations [1]. The failure to control frequency oscillation can cause severe damage to power system stability and reliability. In the worst case, it can cause large-scale power outages and even blackouts. There have been several incidents of low-frequency inter-area oscillation, such as the one on 14 August 2003 at the Eastern Interconnection located in the United States [2]. This incident caused 45 million people to lose their power supply for periods of up to three hours. The outage was caused by poor damping of low-frequency oscillations. Another incident took place in the southern region, where the power system broke down on 15 September 2011. The incident was a result of poor frequency oscillation damping of the power system. Such events are harder to avoid with traditional solutions because the maximum available transfer capability is limited [3–5]. For example, traditional power engineering approaches damp oscillations with power system stabilizers (PSSs), relying on local measurements. However, it is hard to satisfy controllability and observability over inter-area modes if signals are measured locally [6]. Fortunately, the observability of

inter-area modes improved with the advent of wide-area measurement systems (WAMS) and the implementation of phasor measurement units (PMUs) [7–11]. Due to the communication delay in modern Information and Communication Technologies (ICT) [12], the uncertainty of the communication delay becomes an essential research focus in power system damping control.

Since the communication delay significantly affects the damping control performance, researchers have proposed various control methods to solve this issue [13–15]. Mokhtari et al. develop a WAMS controller using fuzzy logic algorithms [16]. The purpose is to mitigate the adverse impact stemming from the continuous latency in the inter-area mode. Meanwhile, a series of inter-area oscillation damping controllers were developed to handle the varying-latency issues [17,18]. Ref. [19] utilizes a special controller to reduce the latency below 300 ms. The work in [20] employs networked predictive control to mitigate inter-area oscillations in power systems. While [21] utilizes $H\infty$ control to achieve effective damping.

Furthermore, ref. [22] proposes a multi-input multi-output (MIMO) ARMAX model. It has similar accuracy as the MIMO subspace state space model but a lower model order. From a device perspective, the static VAR compensators are adopted under various system conditions and types of generations [23–25]. Although these methods have merits from a different perspective, some of them neglect communication delays and some assume correct network topology and system parameters. Unfortunately, such assumptions are hard to achieve in reality, due to their lack of accessibility, network growth, and the instantaneous communication congestion condition.

To target these issues, intelligent methods that are physically model-free are proposed. For example, ref. [26] takes advantage of the data management technique CART, while ref. [27] utilizes a data-driven control method for damping control. Ref. [28] uses a deep learning WADC, but such a method relies extensively on past data and is unable to adapt to the changes in transmission networks. One observation shows that the exploration of the system does help in capturing such transformations. Fortunately, reinforcement learning (RL) provides a platform to understand the environment and learn the control policy accordingly. Among different RL methods, Q-learning can handle problems with stochastic transitions and rewards using a value function. Ref. [29] leverages this capability of learning a stochastic control through exploring the network by using Q-learning. However, Q-learning fails where there is a large state space [30–33].

The contributions of this paper are summarized here. First, to overcome control challenges of large state space, we can combine the advantage of RL and deep learning to provide a stochastic and robust control through WAMS, so that both the uncertainties and time-varying delays can be taken into account through interactive learning. Second, this paper introduces a novel policy-based RL method to address the damping control due to unknown latency in inter-area oscillation. Specifically, we build a power system testbed for RL's interactive environment. Third, we define the state and action most suitable for damping control. In the end, a reward function that considers the physical measurements and sustainability of RL is proposed.

This paper has the following outlines: Section 2 explains the background knowledge of the RL and policy gradient method. Section 3 elaborates on the specific design of the RL-based controller, including the design of state, action, and reward to maximize the control benefits and merge them into the power system concept. Further simulation results and shortcomings of the policy-based RL method are described in Section 4, followed by the conclusions in Section 5.

2. Background

2.1. The Principles of RL Algorithms

Power systems damping control deals with uncertainties and ambiguities in the entire power system. RL is a perfect tool to solve such issues because it helps the control agents learn the optimal control actions with the highest accumulative reward. Essentially, most RL

algorithms explore the environment by viewing it as a Markov decision process (MDP) [34]. The MDP usually has a tuple expression (S, A, E, r). It is composed of a state space S, an action space A, and a transition function $E[S_{t+1}|s_t, a_t]$ that predicts the next step s_{t+1} based on the current state and action (s_t, a_t). Each of these (s_t, a_t) is used to calculate a reward at the present state while taking the present action. In power systems, the state-action pair is defined as the control action taken under present operating conditions, whereas the reward is the score obtained after a control action. The policy function π reflects the agent's "intelligence" since it decides the action to be taken in some state s_t. Therefore, the objective of the RL agent is to formulate an optimal model for policy-making. This model could use a gradient-based learning rule for the policy to achieve the highest accumulative reward. The corresponding RL operational principle diagram shown in Figure 1, clearly displays the relationship between environment and agent. The agent acquires the state of the system under study through measurement and communication devices. Based on the observations, the agent then determines the corresponding action to control the state of the environment through the calculation of the reward. To improve control performance, the agent updates its policy at each time step of the process.

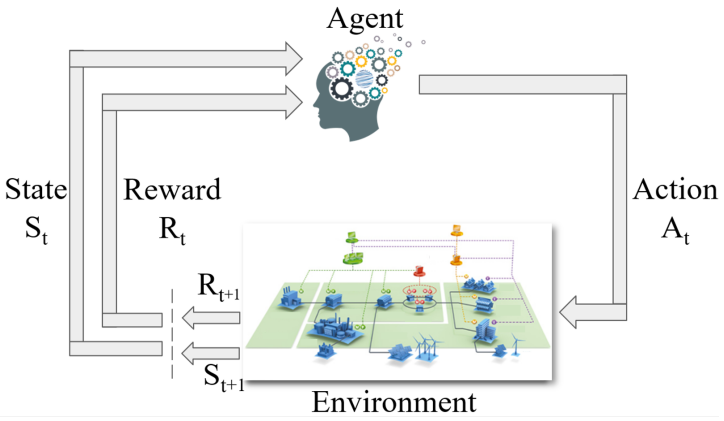

Figure 1. The agent-environment interaction in reinforcement learning.

2.2. DDPG Algorithm

Many researchers focus on using machine learning methods such as deep Q-network (DQN) to address latency issues in modern controls. Despite the fact that the DQN overcomes the exploring challenges in high-dimensional space, it cannot work in continuous action space. This is not acceptable in power system damping control where the action space is continuous. Therefore, the DDPG algorithm seems to be a promising solution because it relies on an actor-critic model to explore the continuous action space [35]. Besides, DDPG has good accuracy in learning under complex environments as shown in [36]. Its value function is expressed as follows:

$$Q(s_t, a_t) = E[r(s_t, a_t) + \gamma \max_a E[Q(s_{t+1}, a_{t+1})]. \quad (1)$$

In the critical networks, we use Q to estimate the state of the WAMS and follow a specific distribution [35], as it estimates the effectiveness of the action being taken. In power system damping control, the action can be the reference value for the generator terminal voltage. To predict the actions, an actor-network $\mu(s_{t+1})$ is used, as it samples the states and conducts the value function estimation. However, the estimation is supervised by the critic and its evaluation is denoted as:

$$y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'}). \quad (2)$$

The above two equations are approximated by neural networks and parameterized by θ^Q and θ^μ [35]. In power systems, they are the parameters for the control performance models. For M time steps, we use the following loss function:

$$\text{Loss} = \frac{1}{M}\sum_{i=1}^{M}(y_i - Q(s_i, a_i))^2, \tag{3}$$

where i is the mini-batch sample number [35]. The actor-network is iteratively updated through the chain rule based on the initial reward distribution. The expected return is parameterized by θ and expressed as:

$$\nabla_{\theta_\mu} J \approx \frac{1}{M}\nabla_a Q(s_i, a_i|\theta^Q)\nabla_{\theta_\mu}\mu(s_i|\theta^\mu). \tag{4}$$

The target actor and critic are denoted by Q' and μ'. Starting with their initial values, they are updated respectively in an iterative way [35].

3. Reinforcement Learning Based Controller

Based on the DDPG algorithm, we further realized that the twin-delayed deep deterministic policy gradient algorithm is needed to avoid the overestimation bias of DDPG. The framework of the proposed damping control scheme is shown in Figure 2, and is motivated by WAMS, in which the PMUs transmit phasor data to the phasor data concentrator (PDC). Since different media are deployed, the transmission delay affects the receiving time of the important information for determining the algorithm inputs. The goal of Figure 2 is to achieve state observability of the whole system for RL control.

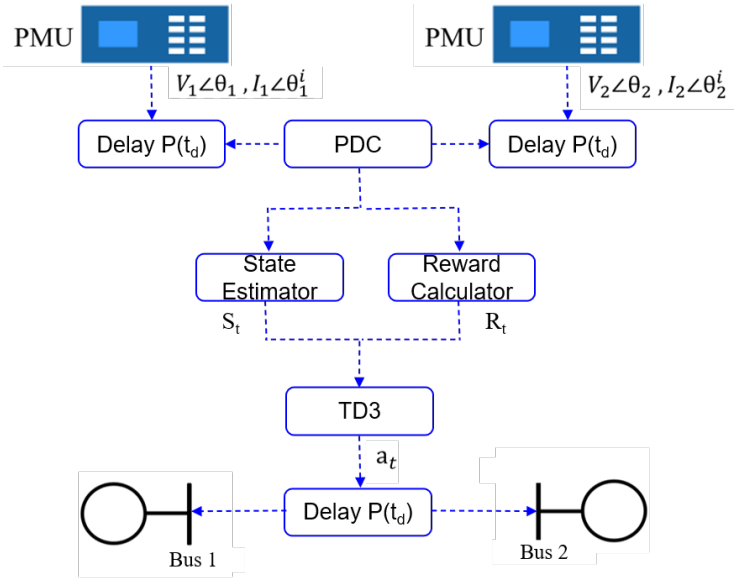

Figure 2. Framework of the overall scheme. The communication links are exhibited by blue dotted lines.

The twin-delayed deep deterministic policy gradient controller interprets the PDC data and calculates the states and rewards. After that, control actions are output through the learning process. The neural networks are used to represent the actor and the critic. The backpropagation of the networks is realized by minimizing the loss function, as shown in Figure 3. To mitigate the oscillations and prevent the overestimation of Q-values, we use a pair of actors and critics to form a twin-delayed deep deterministic policy gradient algorithm. However, further work is needed for defining the states and the actions. One of the challenges is the twin-delayed deep deterministic policy gradient reward calculation,

which is the key to the RL controller design. In the following subsections, we demonstrate how the RL controller is designed.

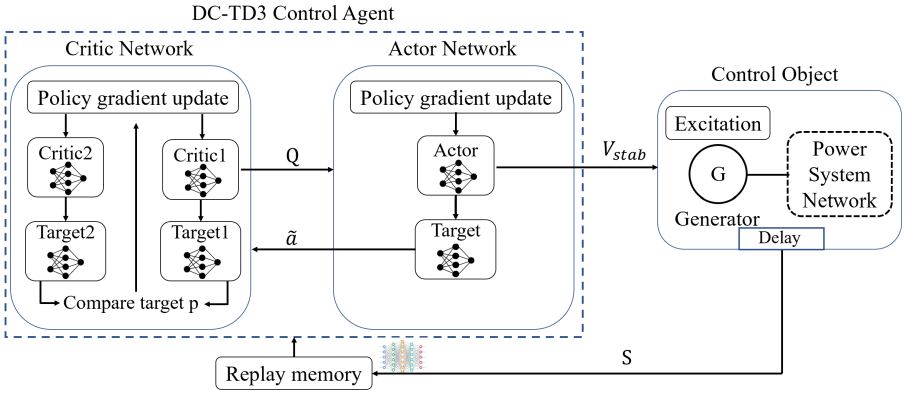

Figure 3. Internal structure and the implementation diagram of the twin-delayed deep deterministic policy gradient controller.

3.1. State of the Controller

The twin-delayed deep deterministic policy gradient controller has three elements: the state, the action, and the reward. We start with the design of the states. The generator voltage, current, and phase angle are monitored through PMUs [37]. We define the states s_t for the generators under study, denoted by $g = 1, \cdots, G$. Then, we define $\omega_{t,g}$ as the generator speed. The associated speed perturbation is defined as $\Delta\omega_{t,g}$. We use $\theta_{t,b}$ for the generator phase angle. In addition, the bus number and time are denoted by b and t, with an upper bound of B and T, respectively. Simply, the speed perturbation is:

$$\Delta\omega_{t,g} = \omega_{t,g} - \omega_{t-1,g}. \tag{5}$$

Meanwhile, the states are summarized as follows:

$$\begin{aligned} s_{t,1} &= \{\omega_{t,1}, \omega_{t,2}, \omega_{t,3}, \cdots, \omega_{t,G}\}, \\ s_{t,2} &= \{\Delta\omega_{t,1}, \Delta\omega_{t,2}, \Delta\omega_{t,3}, \cdots, \Delta\omega_{t,G}\}, \\ s_{t,3} &= \{\theta_{t,1}, \theta_{t,2}, \theta_{t,3}, \cdots, \theta_{t,B}\}, \\ s_T &= s_{t,1} \cup s_{t,2} \cup s_{t,3}. \end{aligned} \tag{6}$$

We design the state to directly capture the rotor speed; therefore, we include $s_{t,1}$. Meanwhile, to capture the dynamic characteristics of the rotor, we utilize the rotor speed deviation to monitor the direct results after an action. Therefore, we have $s_{t,2}$ in Equation (6). Since the voltage angle is another significant factor that quantifies the state of the generators, we incorporate $s_{t,3}$ in the design of the state.

3.2. Action of the Controller

After designing the controller's state, it is important to identify the control action. For the twin-delayed deep deterministic policy gradient controller, it serves as a special power system stabilizer (PSS) that regulates the synchronous generator g's field winding voltage ($v_{t,g}$) at time t. Consequently, the twin-delayed deep deterministic policy gradient controller's action is the increasing and decreasing of field voltages for generators. Since the voltage is established and controlled immediately, we simply use the resulting voltage as the action. When there are multiple generators, the voltage control actions are grouped in an action vector:

$$a_t = \{v_{t,1}, v_{t,2}, v_{t,3}, \cdots, v_{t,G}\}. \tag{7}$$

By grouping the voltage signals, the learning agent can easily process and deliver the control signal to the automatic voltage regulator of each generator.

3.3. Reward Design for Enhanced Control Results

Now, we create the reward function. To dampen the power systems' frequency oscillations, different indicators are collected, including the rotor speed, its deviation, and phase angle deviation between remote buses. To solve the high dimensionality and complexity of stability problems, we not only use variables for the power system but add more control effort from RL into the reward function to improve the performance. The form of reward function is shown in the following:

$$\begin{aligned} r_t = & -c_1 \sum_{g=1}^{G} (\omega_{t,g})^2 - c_2 \sum_{g=1}^{G} (\Delta\omega_{t,g})^2 \\ & - c_3 \sum_{\substack{i,j \in B \\ i \neq j, i<j}}^{B} (\theta_{t,i} - \theta_{t,j})^2 \\ & - c_4 \frac{T_s}{T_f} - c_5 \sum_{g=1}^{G} (v_{t-1,g})^2, \end{aligned} \tag{8}$$

where, v_{t-1} is the variable of the action vector. It keeps a record of the feedback from the previous time steps. This variable will help have higher reward values and reduce the system oscillations. It is observed that there are five terms in Equation (8), ranging from physical quantity associated reward (the first three terms) to the episode control (the fourth term) and the feedback of actions (the fifth term). The first term gives the ability to control the speed of the generators $\omega_{t,g}$. The second term captures the generator speed variation. The third term measures the bus angle differences for the generators under study. Due to WAMS, we are able to obtain information that describes the angle difference in a large area. This provides a fresh perspective for better observation. The fourth term refers to the constant reward for preventing the termination of the episode due to zero reward. The fifth term in (8) refers to the feedback for the action spaces from previous time steps. As we capture the major variables that impact the system stability, we add them into one reward function. To emphasize the feedback, the fifth term is designed in a square form. The reward design is novel, since it quantifies the physical values, includes the episode control, and adds the feedback from the RL agent's action. Together, these innovations help to achieve superior control performance.

3.4. Twin-Delayed Deep Deterministic Policy Gradient Method

Conventional solutions for damping control are usually model-based. However, the parameter variation over time, and the communication latency, are major issues for model-based solutions. Under this circumstance, RL algorithms are gaining popularity. However, some RL algorithms, like Q-learning, suffer from the overestimation issue in the Q-values. Therefore, we solve the overestimation issue by implementing the TD3 algorithm [38] to the power system damping control. The twin-delayed deep deterministic policy gradient algorithm is an off-policy RL method that uses deep neural networks to compute an optimal policy that damps down the power system oscillations.

In the RL-based controller design, we expect the RL-based controller to have some salient features. First, the control agent manages and stores the actor and critic networks, and both continue to improve the stability during the learning process. The twin-delayed deep deterministic policy gradient algorithm maintains the actor and critic networks. The actor network inputs the system state and outputs the damping control action. The mapping from the input to the output is a control policy, learned in the twin-delayed deep deterministic policy gradient algorithm; whereas the critic evaluates the action made by the actor.

Next, the twin-delayed deep deterministic policy gradient controller maintains a dynamic replay memory collected from an environment interface. The reply memory stores the "experience" about the environment's reaction to the control agent's action.

The replay memory is utilized so that it mitigates the association in the sampled data by randomly selecting a batch of data from the replay memory at each time step. To dampen oscillations of highly non-linear power systems under communication delays, we design the twin-delayed deep deterministic policy gradient controller with two deep neural networks for the actor and the critic respectively.

Lastly, for the loss function design of the twin-delayed deep deterministic policy gradient controller, we use the deterministic policy gradient method for the actor training:

$$\nabla_{\theta_\mu} L \approx \frac{1}{M} \nabla_a Q(s_i, a_i | \theta^Q) \nabla_{\theta_\mu} \mu(s_i | \theta^\mu). \tag{9}$$

3.5. Avoid the Overestimation Bias

To make sure the control tasks are in a continuous action space, an actor-critic setting is adopted. To avoid high bias in the damping control, the clipped Double Q-learning method is used. This method takes advantage of the overestimation bias and sets it as the upper limit of the real estimated value [38]. This method is inspired by Double DQN [39], where the target network is used to approximate the value function and extract the policy. When translated into the actor-critic environment, we update the present policy instead of the target policy with two actor values ($\pi_{\phi_1}, \pi_{\phi_2}$), two critic values ($Q_{\theta_1}, Q_{\theta_2}$), and the objective p:

$$p_1 = r + \gamma Q_{\theta_2'}(s', \pi_{\phi_1}(s')), \tag{10}$$

$$p_2 = r + \gamma Q_{\theta_1'}(s', \pi_{\phi_2}(s')). \tag{11}$$

To prevent the propagation of the overestimation when the smaller Q_θ has already overestimated the true value, the strategy is to use the biased Q_θ value as the upper bound of the less biased one [38]. After summing up the experience reward r and the minimized reward of the critics, we formulate the value function target as follows:

$$p_1 = r + \gamma \min_{i=1,2} Q_{\theta_i'}(s', \pi_{\phi_1}(s')). \tag{12}$$

4. Numerical Validation

In this section, first we will discuss the validation setup. Then, we demonstrate the performance of the proposed control agent with validation on the latency. Lastly, a detailed comparison with the DDPG and a classical damping control method is shown.

4.1. Benchmark System

Various simulation studies were performed in benchmark systems like the Kundur system and the IEEE 39-bus system. The performances are similar, but due to the space limit, we use the Kundur's system in Figure 4 for illustration, as it is a widely used benchmark system for dynamic oscillation studies [40]. There are two areas in the system. Area 1 includes generators G1 and G2, while area 2 has generators G3 and G4, as shown in Figure 4. Both areas 1 and 2 share two identical generators, 20 kV, 900 MW, except for the inertia value. The two areas are connected through two 230 kV transmission lines. The distance between the two areas is 220 km; in Figure 4, this is the distance between bus 7 and 9. Table 1 summarizes the benchmark system parameters. There are two solar farms connected to bus 7 and 9, respectively. The benchmark system presents a stressed operating condition, where area 2 imports 413 MW from area 1.

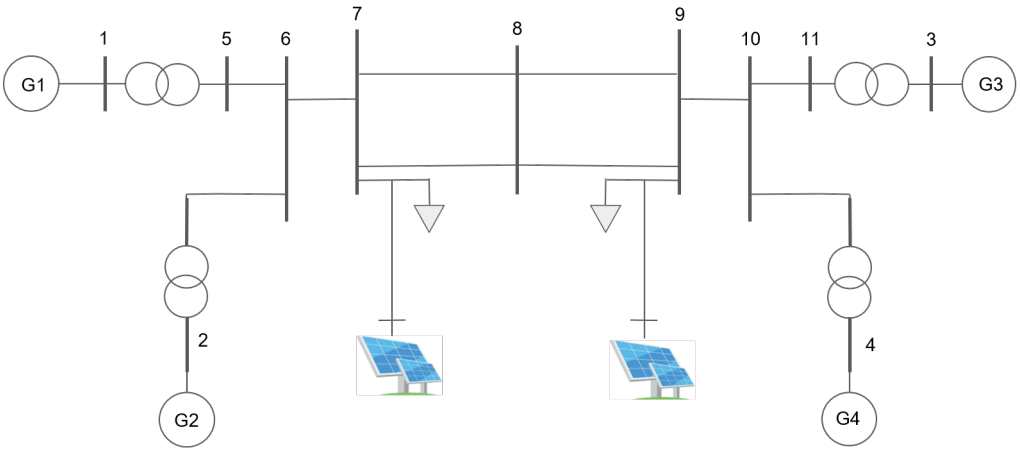

Figure 4. This modified Kundur system has four synchronous machines and two aggregated residential PV systems.

Table 1. Parameters of the benchmark system under study.

Name	Value
Generator	20 kV/900 MVA
Synchronous machine inertia	6.5 s, 6.175 s
Thermal plant exciter gain	200
Solar capacity (PV1, PV2)	100 MW
Surge impedance loading	140 MW
Area 2 power generation	700 MW

4.2. Twin-Delayed Deep Deterministic Policy Gradient Control Agent: Fast Learning Curve

To show the robustness of the damping controller, the controller is comprehensively trained under various communication latency. Using the proposed controller design in Section 3, we obtain the learning results shown in Figure 5. The average reward reflects the control performance directly. We also notice that attempts at the initial stage often vary and that the average reward gradually increases at a later stage. As observed in Figure 5, after 100 episodes of simulation, the episode reward of the agent is shown to reach a value close to zero—the highest value in the whole learning curve.

Being out of synchronization can result in system collapse, so we determine a low reward for this scenario. When the system loses synchronization in an episode, the agent learns associated parameters so that the system does not explore the outage of the system and learns what parameters to avoid next time. Therefore, we design such cases as an indicator to terminate the episode training in the learning process. We observe that the agent gradually increases to an appropriate policy that reflects proper control strategies when the episode number is below 40, as shown in Figure 5. As the attempts increase, a high average reward is established. Episodes 40 and 85 present the convergence points for the average reward and episode reward. As the learning process accumulates, extensive exploration leads to a stabilization stage as shown in Figure 5. The control agent then becomes very effective in damping the oscillation.

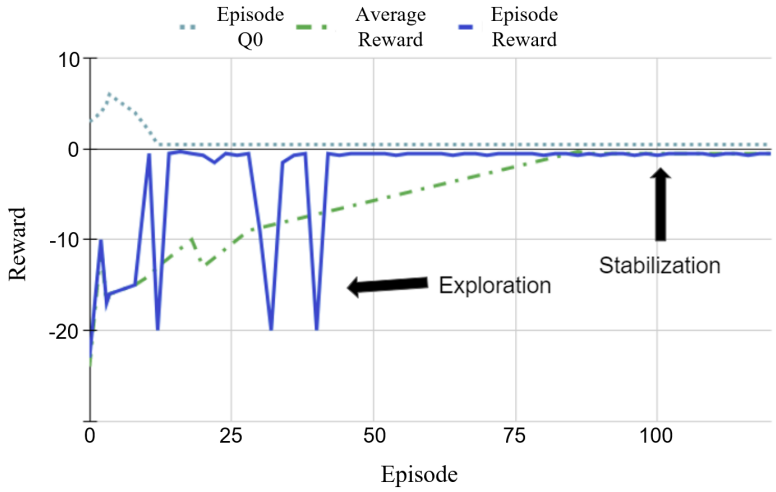

Figure 5. This typical learning curve shows the x and y axis, which are the number of episodes and the rewards, respectively. The blue line indicates the episode reward for each episode. The red line reflects the average reward (a running average reward value), while the yellow line represents the episode Q0 (the reward computed as the beginning of each episode). See Table 2 for the parameter details.

The advantage of this reward function design is presented in Figure 6, which clearly shows the necessity of the five terms in the reward equation. When there are less than three terms in the reward function, the learning curve cannot be accomplished. The first three terms in Equation (8) show that the green learning curve in Figure 6 is much worse than the blue curve that has five terms. The proposed twin-delayed deep deterministic policy gradient technique with the five-term reward function gains higher reward values in general, which indicates the system becomes the stabilization condition faster than regular TD3 with only three terms.

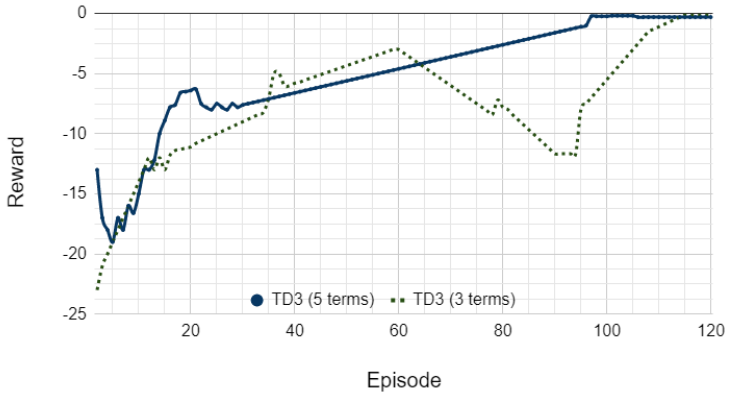

Figure 6. This result comparison uses the proposed twin-delayed deep deterministic policy gradient algorithm under a 3-term and 5-term reward function, where 3 terms refer to the first three terms in Equation (8), and 5 terms refer to the entire Equation (8). The reward shown here refers to the running average reward value.

4.3. Control Agent: Robust to Communication Latency

We analyze the control performance by altering the average communication delay in the system. The communication latency depends on the communication medium. For example, fiber optic cables have a one-way delay of 100–150 ms, whereas satellite links exhibit a latency range of 500–700 ms. A full list of the latency ranges can be found in [41]. Based on these ranges, four scenarios (S1–S4) that fully capture the delay range are created between 0.13∼0.19 s. The details of the four testing scenarios are shown in Appendix A.

Since the signal routing and its internal mechanism present uncertainty and variability, the controller learns these characteristics in each episode. Figure 7 shows that the proposed agent with the proposed controller has an excellent performance in controlling the generators, but shows that the performance based on the variation of average communication delay successfully damps down the oscillations. To show the improvement of the proposed control scheme, we compare the proposed method with the existing control scheme—the multi-band power system stabilizer (MBPSS). MBPSS is one of the classical practices in IEEE Std 421.5 [42]. We test the performance of MBPSS under the same system operating conditions as the proposed method. In steady-state conditions, the control difference is subtle. However, under transient conditions, the proposed twin-delayed deep deterministic policy gradient method show faster control performance than MBPSS. As shown in Figure 7, MBPSS, with a communication latency of 0.13 s, brings the machine speed to normal with almost 20 s under the test scenario. This is almost twice the time as the proposed method.

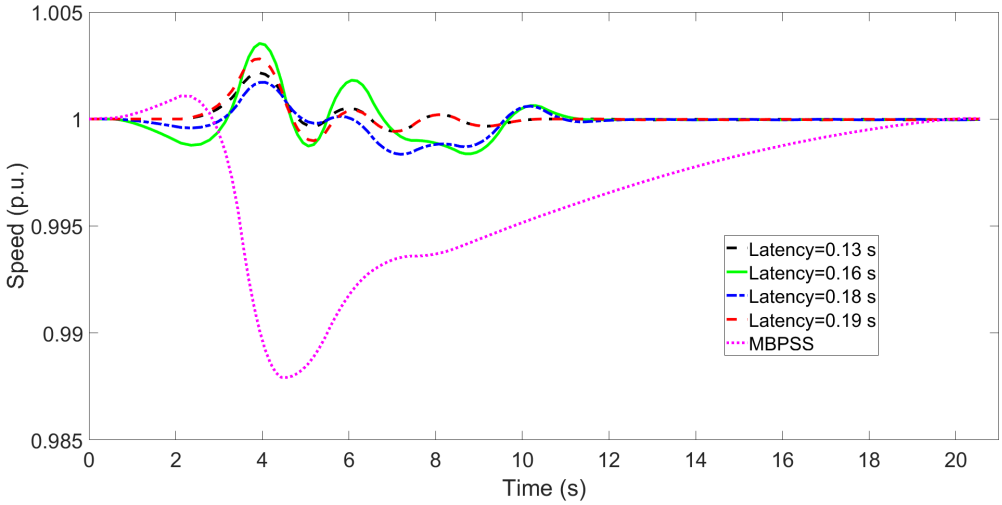

Figure 7. The control performance is based on the variation of average communication delay. Due to the stochastic nature of the proposed RL algorithm, the damping control results are different under various latency. This figure represents the results selected among extensive simulations; however, the largest speed deviation is always well contained in all test cases.

As discussed in the literature, conventional methods suffer from the latency issue during their control efforts. Through multiple simulations, we observe that the performance of the proposed control agent is no longer dependent on the communication delay due to the fact that the largest time latency is not always associated with the highest speed deviation. Figure 7 shows one of the selected control results, where, unlike conventional methods, the shortest time delay does not result in the least amount of speed deviation. Here, the yellow line with 0.18 s of time delay presents the smallest speed deviation. The results show that the proposed RL agent decouples its performance with the time latency in it's communication. It is the hyper-parameters in the RL algorithm that affect the damping performance.

4.4. Performance Comparison with DDPG

The compared results between the twin-delayed deep deterministic policy gradient and DDPG agents are demonstrated in Figure 8, including all the agent discount factors and batch sizes in Table 2. The twin-delayed deep deterministic policy gradient agent achieves the optimal policy faster than the DDPG agent does; whereas, the DDPG agent could not converge into the stabilization condition in limited episodes. However, it has only several explorations before reaching out to the optimal policy. In other words, there could be the case that the agent might not learn the parameters.

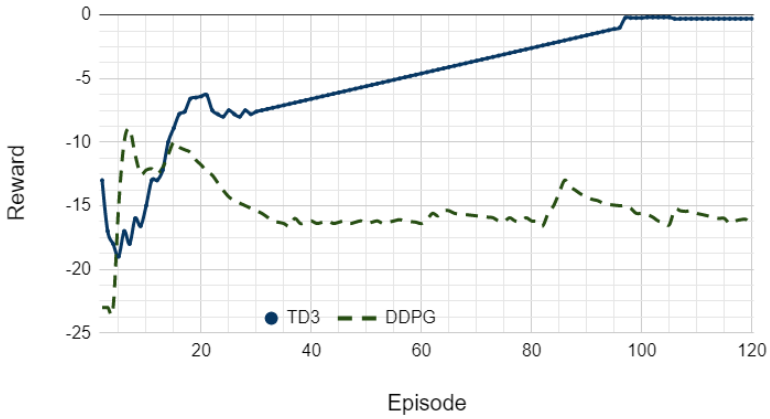

Figure 8. Result comparison between the proposed and DDPG agents. The parameters of both agents are presented in Table 2.

Table 2. Parameters of twin-delayed deep deterministic policy gradient and DDPG agents.

Parameter Type	Value
Number of input states	12
Number of output actions	4
Discount factor γ	0.75
Sample time (sec)	0.1
Replay memory	500,000

5. Conclusions

We solve problems of instability in large systems by proposing a twin-delayed deep deterministic policy gradient control agent that takes into consideration all the uncertainties of the unbalanced systems. In this way, the system is optimally explored to damping down frequency oscillations while keeping the system's balance within defined limits. Besides, the novel design of the state, action, and the reward is described in detail. By using the twin-delayed deep deterministic policy gradient algorithm, we show that low-frequency oscillation can be significantly improved from its learning curves. The simulation results show the proposed controller has a fast convergence rate and is robust to communication latency variation. When compared to other conventional damping control methods, the proposed twin-delayed deep deterministic policy gradient algorithm most effectively dampen the speed oscillation. Future work can focus on the knowledge transfer of the learned control experience. Specifically, if the control agent does not need to acquire the control knowledge from the scratch, the proposed twin-delayed deep deterministic policy gradient method would show great advantages of generality when applied to various power systems.

Author Contributions: Conceptualization, Q.C., G.K. and Y.W.; methodology, G.K. and Y.W.; validation, Q.C. and G.K.; formal analysis, Q.C.; investigation, G.K.; resources, Y.W.; writing—original draft preparation, Q.C.; writing—review and editing, Y.W.; visualization, G.K.; supervision, Y.W.; project administration, Q.C.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Collaborative Research of Learning and Optimizing Power Systems: A Geometric Approach, Award Number: 1810537, as well as the CAREER award of Faithful, Reducible, and Invertible Learning in Distribution System for Power Flow, Award Number: 2048288.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. The Parameters of Four Testing Scenarios

The four testing scenarios that were used in Section 4 are listed here. They include four representative latency cases under different means and variances of the signals.

Table A1. Mean and variance communication delay using the Gaussian distributed random signal.

	Mean (ms)	Variance
S1	130	0.195
S2	160	0.024
S3	180	0.027
S4	190	0.0285

References

1. Yohanandhan, R.V.; Srinivasan, L. Decentralized Wide-Area Neural Network Predictive Damping Controller for a Large-scale Power System. In Proceedings of the IEEE International Conference on Power Electronics, Drives and Energy Systems, Chennai, India, 18–21 December 2018; pp. 1–6.
2. Andersson, G.; Donalek, P.; Farmer, R.; Hatziargyriou, N.; Kamwa, I.; Kundur, P.; Martins, N.; Paserba, J.; Pourbeik, P.; Sanchez-Gasca, J.; et al Causes of the 2003 major grid blackouts in North America and Europe, and recommended means to improve system dynamic performance. *IEEE Trans. Power Syst.* **2005**, *20*, 1922–1928. [CrossRef]
3. Azad, S.P.; Iravani, R.; Tate, J.E. Damping inter-area oscillations based on a model predictive control HVDC supplementary controller. *IEEE Trans. Power Syst.* **2013**, *28*, 3174–3183. [CrossRef]
4. Klein, M.; Rogers, G.J.; Kundur, P. A fundamental study of inter-area oscillations in power systems. *IEEE Trans. Power Syst.* **1991**, *6*, 914–921. [CrossRef]
5. Zenelis, I.; Wang, X. Wide-Area Damping Control for Interarea Oscillations in Power Grids Based on PMU Measurements. *IEEE Control Syst. Lett.* **2018**, *2*, 719–724. [CrossRef]
6. Aboul-Ela, M.E.; Sallam, A.A.; McCalley, J.D.; Fouad, A.A. Damping controller design for power system oscillations using global signals. *IEEE Trans. Power Syst.* **1996**, *11*, 767–773. [CrossRef]
7. Zhang, S.; Vittal, V. Design of wide-area power system damping controllers resilient to communication failures. *IEEE Trans. Power Syst.* **2013**, *28*, 4292–4300. [CrossRef]
8. Kamwa, I.; Grondin, R.; Hebert, Y. Wide-area measurement based stabilizing control of large power systems—A decentralized/hierarchical approach. *IEEE Trans. Power Syst.* **2001**, *16*, 136–153. [CrossRef]
9. Ma, J.; Wang, T.; Wang, Z.; Thorp, J.S. Adaptive Damping Control of Inter-Area Oscillations Based on Federated Kalman Filter Using Wide Area Signals. *IEEE Trans. Power Syst.* **2013**, *28*, 1627–1635. [CrossRef]
10. Erlich, I.; Hashmani, A.; Shewarega, F. Selective damping of inter area oscillations using phasor measurement unit signals. In Proceedings of the IEEE Trondheim PowerTech, Trondheim, Norway, 19–23 June 2011; pp. 1–6.
11. Hashmy, Y.; Yu, Z.; Shi, D.; Weng, Y. Wide-area measurement system-based low frequency oscillation damping control through reinforcement learning. *IEEE Trans. Smart Grid* **2020**, *11*, 5072–5083. [CrossRef]
12. Vu, T.L.; Turitsyn, K. Lyapunov Functions Family Approach to Transient Stability Assessment. *IEEE Trans. Power Syst.* **2016**, *31*, 1269–1277. [CrossRef]
13. Roy, S.; Patel, A.; Kar, I.N. Analysis and Design of a Wide-Area Damping Controller for Inter-Area Oscillation with Artificially Induced Time Delay. *IEEE Trans. Smart Grid* **2019**, *10*, 3654–3663. [CrossRef]
14. Bento, M.E.C. Fixed Low-Order Wide-Area Damping Controller Considering Time Delays and Power System Operation Uncertainties. *IEEE Trans. Power Syst.* **2020**, *35*, 3918–3926. [CrossRef]
15. Lu, C.; Zhang, X.; Wang, X.; Han, Y. Mathematical Expectation Modeling of Wide-Area Controlled Power Systems with Stochastic Time Delay. *IEEE Trans. Smart Grid* **2015**, *6*, 1511–1519. [CrossRef]
16. Mokhtari, M.; Aminifar, F.; Nazarpour, D.; Golshannavaz, S. Wide-area power oscillation damping with a fuzzy controller compensating the continuous communication delays. *IEEE Trans. Power Syst.* **2013**, *28*, 1997–2005. [CrossRef]
17. Surinkaew, T.; Ngamroo, I. Inter-Area Oscillation Damping Control Design Considering Impact of Variable Latencies. *IEEE Trans. Power Syst.* **2019**, *34*, 481–493. [CrossRef]
18. Yu, S.S.; Chau, T.K.; Fernando, T.; Iu, H.H.C. An Enhanced Adaptive Phasor Power Oscillation Damping Approach with Latency Compensation for Modern Power Systems. *IEEE Trans. Power Syst.* **2018**, *33*, 4285–4296. [CrossRef]
19. Roberson, D.; O'Brien, J.F. Variable Loop Gain Using Excessive Regeneration Detection for a Delayed Wide-Area Control System. *IEEE Trans. Smart Grid* **2018**, *9*, 6623–6632. [CrossRef]

20. Yao, W.; Jiang, L.; Wen, J.; Wu, Q.H.; Cheng, S. Wide-Area Damping Controller of FACTS Devices for Inter-Area Oscillations Considering Communication Time Delays. *IEEE Trans. Power Syst.* **2014**, *29*, 318–329. [CrossRef]
21. Li, M.; Chen, Y. A Wide-Area Dynamic Damping Controller Based on Robust H_∞ Control for Wide-Area Power Systems with Random Delay and Packet Dropout. *IEEE Trans. Power Syst.* **2018**, *33*, 4026–4037. [CrossRef]
22. Liu, H.; Zhu, L.; Pan, Z.; Bai, F.; Liu, Y.; Liu, Y.; Patel, M.; Farantatos, E.; Bhatt, N. ARMAX-Based Transfer Function Model Identification Using Wide-Area Measurement for Adaptive and Coordinated Damping Control. *IEEE Trans. Smart Grid* **2017**, *8*, 1105–1115. [CrossRef]
23. Vahidnia, A.; Ledwich, G.; Palmer, E.W. Transient Stability Improvement through Wide-Area Controlled SVCs. *IEEE Trans. Power Syst.* **2016**, *31*, 3082–3089. [CrossRef]
24. Bian, X.Y.; Geng, Y.; Lo, K.L.; Fu, Y.; Zhou, Q.B. Coordination of PSSs and SVC Damping Controller to Improve Probabilistic Small-Signal Stability of Power System with Wind Farm Integration. *IEEE Trans. Power Syst.* **2016**, *31*, 2371–2382. [CrossRef]
25. Zhang, K.; Shi, Z.; Huang, Y.; Qiu, C.; Yang, S. SVC damping controller design based on novel modified fruit fly optimisation algorithm. *IET Renew. Power Gener.* **2018**, *12*, 90–97. [CrossRef]
26. Lala, J.A.O.; Gallardo, C.F. Adaptive Tuning of Power System Stabilizer Using a Damping Control Strategy Considering Stochastic Time Delay. *IEEE Access* **2020**, *8*, 124254–124264. [CrossRef]
27. Shi, X.; Cao, Y.; Shahidehpour, M.; Li, Y.; Wu, X.; Li, Z. Data-Driven Wide-Area Model-Free Adaptive Damping Control with Communication Delays for Wind Farm. *IEEE Trans. Smart Grid* **2020**, *11*, 5062–5071. [CrossRef]
28. Jhang, S.; Lee, H.; Kim, C.; Song, C.; Yu, W. ANN Control for Damping Low-frequency Oscillation using Deep learning. In Proceedings of the IEEE Australasian Universities Power Engineering Conference, Auckland, New Zealand, 27–30 November 2018; pp. 1–4.
29. Duan, J.; Xu, H.; Liu, W. Q-Learning-Based Damping Control of Wide-Area Power Systems Under Cyber Uncertainties. *IEEE Trans. Smart Grid* **2018**, *9*, 6408–6418. [CrossRef]
30. Araghi, S.; Khosravi, A.; Johnstone, M.; Creighton, D.C. A novel modular Q-learning architecture to improve performance under incomplete learning in a grid soccer game. *Eng. Appl. Artif. Intell.* **2013**, *26*, 2164–2171. [CrossRef]
31. Das, P.; Behera, D.H.; Panigrahi, B. Intelligent-based multi-robot path planning inspired by improved classical Q-learning and improved particle swarm optimization with perturbed velocity. *Eng. Sci. Technol. Int. J.* **2016**, *19*, 651–669. [CrossRef]
32. Shinohara, S.; Takano, T.; Takase, H.; Kawanaka, H.; Tsuruoka, S. Search Algorithm with Learning Ability for Mario AI—Combination A* Algorithm and Q-Learning. In Proceedings of the Australasian Conference on Information Systems International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Kyoto, Japan, August 2012; pp. 341–344.
33. Cui, Q.; Hashmy, S.M.Y.; Weng, Y.; Dyer, M. Reinforcement Learning Based Recloser Control for Distribution Cables with Degraded Insulation Level. *IEEE Trans. Power Deliv.* **2020**, *36*, 1118–1127. [CrossRef]
34. Szepesvari, C.; Sutton, R.S.; Modayil, J.; Bhatnagar, S. Universal Option Models. In Proceedings of the 28th Annual Conference Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 990–998.
35. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.M.O.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
36. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.; Veness, J.; Bellemare, M.; Graves, A.; Riedmiller, M.; Fidjeland, A.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529. [CrossRef] [PubMed]
37. Simon, L.; Swarup, K.S.; Ravishankar, J. Wide area oscillation damping controller for DFIG using WAMS with delay compensation. *IET Renew. Power Gener.* **2019**, *13*, 128–137. [CrossRef]
38. Fujimoto, S.; Hoof, H.; Meger, D. Addressing function approximation error in actor-critic methods. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1587–1596.
39. Hasselt, H.; Guez, A.; Silver, D. Deep Reinforcement Learning with Double Q-Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 2094–2100.
40. Kamwa, I.; Trudel, G.; Gerin-Lajoie, L. Robust design and coordination of multiple damping controllers using nonlinear constrained optimization. *IEEE Trans. Power Syst.* **2000**, *15*, 1084–1092. [CrossRef]
41. Younis, M.R.; Iravani, R. Wide-area damping control for inter-area oscillations: A comprehensive review. In Proceedings of the IEEE Electrical Power & Energy Conference, Halifax, NS, Canada, 21–23 August 2013; pp. 1–6.
42. Kamwa, I.; Grondin, R.; Trudel, G. IEEE PSS2B versus PSS4B: The limits of performance of modern power system stabilizers. *IEEE Trans. Power Syst.* **2005**, *20*, 903–915. [CrossRef]