

## Article

# Deep-Reinforcement-Learning-Based Two-Timescale Voltage Control for Distribution Systems

Jing Zhang<sup>1</sup>, Yiqi Li<sup>1</sup>, Zhi Wu<sup>1,\*</sup> , Chunyan Rong<sup>2</sup>, Tao Wang<sup>2</sup>, Zhang Zhang<sup>2</sup> and Suyang Zhou<sup>1</sup> 

<sup>1</sup> School of Electrical Engineering, Southeast University, Nanjing 210096, China; jzhang@seu.edu.cn (J.Z.); 220192833@seu.edu.cn (Y.L.); suyang.zhou@seu.edu.cn (S.Z.)

<sup>2</sup> Institute of State Grid Hebei Electric Power Company Economic and Technological Research, Shijiazhuang 050000, China; Rongchunyan@hotmail.com (C.R.); Wangtao037@hotmail.com (T.W.); zhzh019@hotmail.com (Z.Z.)

\* Correspondence: zwu@seu.edu.cn; Tel.: +86-150-0518-4780

**Abstract:** Because of the high penetration of renewable energies and the installation of new control devices, modern distribution networks are faced with voltage regulation challenges. Recently, the rapid development of artificial intelligence technology has introduced new solutions for optimal control problems with high dimensions and dynamics. In this paper, a deep reinforcement learning method is proposed to solve the two-timescale optimal voltage control problem. All control variables are assigned to different agents, and discrete variables are solved by a deep Q network (DQN) agent while the continuous variables are solved by a deep deterministic policy gradient (DDPG) agent. All agents are trained simultaneously with specially designed reward aiming at minimizing long-term average voltage deviation. Case study is executed on a modified IEEE-123 bus system, and the results demonstrate that the proposed algorithm has similar or even better performance than the model-based optimal control scheme and has high computational efficiency and competitive potential for online application.

**Keywords:** deep reinforcement learning; two timescales; voltage control; distribution network



**Citation:** Zhang, J.; Li, Y.; Wu, Z.; Rong, C.; Wang, T.; Zhang, Z.; Zhou, S. Deep-Reinforcement-Learning-Based Two-Timescale Voltage Control for Distribution Systems. *Energies* **2021**, *14*, 3540. <https://doi.org/10.3390/en14123540>

Academic Editor: Tek Tjing Lie

Received: 7 May 2021

Accepted: 11 June 2021

Published: 14 June 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

### 1.1. Background and Motivation

The high penetration of distributed generation (DG) energy sources, such as photovoltaic (PV), has made distribution networks faced with the problem of voltage regulation. Usually, the voltage profiles in distribution networks are regulated by the control of slow regulation devices (e.g., on-load tap changers (OLTCs) and shunt capacitors) and fast regulation devices (e.g., PV inverters and static var compensators (SVCs)). While these regulators are all applied to adjust the distribution of reactive power in the grid, the real power flow can also impact the nodal voltages in distribution networks [1,2]. Thus, the real and reactive power control of different devices should be taken into account in order to mitigate possible voltage violations.

The lack of measurement systems (e.g., supervisory control and data acquisition (SCADA) and phasor measurement units (PMUs)) in traditional distribution networks leads to the insufficient measurement of network information, and voltage control methods generally adopt model-based regulation, which rely highly on the precise physical model. In essence, voltage control through real and reactive power optimization is a highly nonlinear programming problem with abundant variables and massive constraints. Solving such problems using mathematical programming methods (e.g., the second-order cone relaxation technique [3] and duality theory [4]) is often limited by the number of variables, and may even fail when the scale of the distribution network is too large. Therefore, heuristic algorithms which are less dependent on the model are applied to solve these problems (e.g., particle swarm optimization (PSO) [5] and genetic algorithm (GA) [6]). However, these

algorithms have the shortcomings of high randomness and long search time, and easily fall into local optimal solutions, and as a result cannot meet the requirement of real-time voltage control in a fast time scale. In addition, in these mathematical programming methods and heuristic algorithms, each optimization solution is independent of each other, and if the actual operating condition (e.g., DG outputs) changes slightly, the previous optimization results are often unable to be made full use of in order to achieve a rapid solution.

Recently, artificial intelligence (AI) technology has developed rapidly and been successfully applied in different fields. Therefore, many scholars are interested in exploring its application in power systems. Among AI technology, as a branch of reinforcement learning (RL) theory, DRL employs the “trial and error” mechanism to interact with the dynamic environment to find the optimal policy for the agent [7]. It has great advantages for solving complex multivariable problems, and has already been employed in various power system optimization problems, such as, electricity market planning [8], multi-agent equilibrium games [9], battery energy arbitrage [10], and scheduling the charging of electric vehicles [11].

The expansion of the coverage of SCADA and PMU [12], as well as the construction of internet of things technologies, has provided an effective way for the advanced applications of voltage control based on data-driven and model-free methods. For example, the authors in [13] employ Q-learning to solve a reactive power optimization problem with the discrete control variables being the transformer tap ratio and shunt capacitor. However, the Q-learning method easily falls into the curse of dimensionality since it applies a table to store the corresponding action-value function, and it is only suitable for the problems where the action space and state space are both discrete. Inspired by the strong exploration capacity of neural networks (NNs) towards high-dimensional searching space, deep Q network (DQN) employs an NN to approximate the action-value function to deal with continuous state domains. In [14], a DQN is used to control shunt capacitors. In [15], a double DQN is applied to achieve the optimal control of thermostatically controlled loads (TCLs) to provide voltage-control-based ancillary services. In [16], a multi-agent DQN-based algorithm is put forward to control switchable capacitors, voltage regulators, and smart inverters, where the continuous variables of voltage regulators and smart inverters are discretized. Further, to deal with the problems with continuous state and action space, deep deterministic policy gradient (DDPG) is put forward, where two NNs are used to approximate the policy function and action-value function. In [17], DDPG is applied to learn the control policy of generators in order to regulate all bus voltages into a predefined safe zone. In [18], a voltage-sensitivity-based DDPG method is proposed to realize the local control of PV inverters, and with a specifically designed reward, the goal of minimizing the network power loss and ensuring safe operation in the grid can be achieved. In [19], a novel adaptive volt-var control algorithm based on DDPG is proposed to realize the real-time control of smart inverters in reactive power management. However, the existing voltage control methods using DRL only focus on the reactive power control, and cannot deal with the discrete and continuous control variables simultaneously.

### *1.2. Novelty and Contribution*

To overcome the above limitations, a DRL method combining a DQN method with DDPG is proposed in this paper to deal with discrete and continuous control variables simultaneously. In this method, the discrete variables are resolved based on a DQN agent and the continuous variables are resolved based on a DDPG agent. Considering the response time and control cost of different equipment, a two-timescale voltage control problem is put forward, where the capacitor configuration is decided in the long timescale and the outputs of PV inverters and energy storage batteries are adjusted in the short timescale. The problem is further turned into the Markov decision process (MDP) and solved by the proposed DRL method. A reward is specially designed to achieve the goal of minimizing the long-term average voltage deviation. Both the DQN agent and DDPG

agent share the same environment (i.e., the distribution network), and are trained by their interaction with the environment.

The contributions of this article are outlined as follows.

- (1) Multiple types of control equipment including capacitors, energy storage batteries, and PV inverters are considered, and a two-timescale voltage control problem is formulated in order to take the control requirements of these devices into account.
- (2) The control variables are assigned to different agents according to their properties, and these agents share the same environment and are trained simultaneously.
- (3) A DRL method is proposed to solve the optimal voltage control problem, where the discrete variables are solved using a DQN agent and the continuous variables are solved using a DDPG agent, and this method can realize real-time control.

## 2. Voltage Problem Formulation

In this section, the two-timescale voltage control problem is formulated, where the control devices, including capacitors, batteries, and PV inverters, are considered.

### 2.1. System Description

In this voltage control problem, the real power regulation is achieved by adjusting the charging/discharging power of batteries, and the reactive power regulation is realized by adjusting the on/off states of capacitors and the outputs of PV inverters. Taking the response time and control cost of different devices into account, the total control process can be divided into long timescale control and short timescale control. Specifically, the entire time period can be divided into  $N_T$  intervals, and each interval can be further divided into  $N_t$  slots. In the long timescale, the capacitors' configurations are made at the beginning of each interval  $T$ , and in the short timescale the outputs of PV inverters and batteries are adjusted at the beginning of each slot  $t$ .

#### (1) Capacitor modeling

During each  $T$ , the reactive power supported by the capacitor installed at bus  $i$ ,  $Q_{Cap,i}(T, t)$ , can be expressed as a function of binary control variable  $a_{cap,i}(T) \in \{0,1\}$ , which indicates the off/on status of the capacitor; that is,

$$Q_{Cap,i}(T, t) = a_{cap,i}(T) \cdot Q_{Cap,i}^{rated} \quad (1)$$

where  $Q_{Cap,i}^{rated}$  is the rated reactive power of the capacitor. When  $a_{cap,i}(T) = 1$ , the capacitor is connected to the grid and the reactive power provided by it is  $Q_{Cap,i}^{rated}$ ; when  $a_{cap,i}(T) = 0$ , the capacitor is disconnected from the grid.

#### (2) Battery modeling

During every  $t$ , the state variable (state of charge) of the battery installed at bus  $i$  is denoted as  $SOC_i(T, t)$ , which is subject to the upper and lower boundaries [20]; that is,

$$SOC_{i,\min} \leq SOC_i(T, t) \leq SOC_{i,\max} \quad (2)$$

The charging/discharging power of the battery,  $P_{Batt,i}(T, t)$ , can be expressed as

$$P_{Batt,i}(T, t) = a_{batt,i}(T, t) \cdot P_{Batt,i}^{max} \quad (3)$$

$$-1 \leq a_{batt,i}(T, t) \leq 1 \quad (4)$$

where  $a_{batt,i}(T, t)$  is the control variable and when  $a_{batt,i}(T, t)$  is positive, the battery is charging. When  $a_{batt,i}(T, t)$  is negative, the battery is discharging. The new state of charge after taking the control action can be represented as

$$SOC_i(T, t+1) = SOC_i(T, t) + P_{Batt,i}(T, t). \quad (5)$$

### (3) PV inverter modeling

Suppose all PV units in the grid are equipped with a smart inverter. On every  $t$ , the active power provided by the PV unit installed at bus  $i$  is known as  $P_{PV,i}(T, t)$  and its apparent power rating is  $S_{PV,i}^{rated}$ . Then, the reactive power provided by the smart inverter,  $Q_{PV,i}(T, t)$ , can be expressed as

$$Q_{PV,i}(T, t) = a_{pv,i}(T, t) \cdot \bar{Q}_{PV,i}(T, t) \quad (6)$$

$$\bar{Q}_{PV,i}(T, t) = \sqrt{\left(S_{PV,i}^{rated}\right)^2 - \left(P_{PV,i}(T, t)\right)^2} \quad (7)$$

$$-1 \leq a_{pv,i}(T, t) \leq 1 \quad (8)$$

where  $a_{pv,i}(T, t)$  is the control variable of the PV inverter.

### 2.2. Two-Timescale Voltage Control Model Formulation

In this paper, a radial distribution network with  $N + 1$  buses is considered, where Bus 0 is the root, representing the point of common coupling. The voltage magnitude, active power, and reactive power are all converted to per unit (p.u.). The objective of the voltage control problem is to minimize the long-term average voltage deviation by configuring the on/off status of capacitors on every interval  $T$  in the long timescale and adjusting the charging/discharging power of batteries and the outputs of PV inverters on every  $t$  in the short timescale. Then, the two-timescale voltage control problem based on power flow equations can be formulated as follows:

$$\min_{\{a_{cap}(T)\}, \{a_{batt}(T, t), a_{pv}(T, t)\}} \mathbb{E} \left[ \sum_{T=1}^{\infty} \sum_{t=1}^{N_t} \sum_{i=1}^N (U_i(T, t) - 1)^2 \right] \quad (9)$$

s.t.

$$(1)-(3) \quad (10)$$

$$\sum_{i \in \psi(j)} \left[ P_{ij}(T, t) - I_{ij}^2(T, t)r_{ij} \right] - P_j(T, t) = \sum_{k \in \phi(j)} P_{jk}(T, t) \quad (11)$$

$$\sum_{i \in \psi(j)} \left[ Q_{ij}(T, t) - I_{ij}^2(T, t)x_{ij} \right] - Q_j(T, t) = \sum_{k \in \phi(j)} Q_{jk}(T, t) \quad (12)$$

$$U_j^2(T, t) = U_i^2(T, t) + \left(r_{ij}^2 + x_{ij}^2\right) I_{ij}^2(T, t) - 2\left(r_{ij}P_{ij}(T, t) + x_{ij}Q_{ij}(T, t)\right) \quad (13)$$

$$I_{ij}^2(T, t) = \left(P_{ij}^2(T, t) + Q_{ij}^2(T, t)\right) / U_i^2 \quad (14)$$

$$P_j(T, t) = P_{L,j}(T, t) + P_{Batt,j}(T, t) - P_{PV,j}(T, t) \quad (15)$$

$$Q_j(T, t) = Q_{L,j}(T, t) - Q_{Cap,j}(T, t) - Q_{PV,j}(T, t) \quad (16)$$

where  $U_j(T, t)$  is the voltage magnitude of bus  $j$ ;  $I_{ij}(T, t)$  is the current amplitude of segment  $(i, j)$ ;  $r_{ij}$  and  $x_{ij}$  are the resistance and reactance of segment  $(i, j)$ , respectively;  $P_{ij}(T, t)$  and  $Q_{ij}(T, t)$  are the active and reactive power flowing from bus  $i$  to bus  $j$ , respectively;  $\psi(j)$  is the parent bus set of bus  $j$ , where the power flows from the parent bus to bus  $j$ ; and  $\phi(j)$  is the child bus set of bus  $j$ , where the power flows from bus  $j$  to the child bus.

It can be observed that the optimization problem (9) involves many control variables, including the continuous variables of batteries and PV inverters and the discrete variables of capacitors, which makes problem (9) non-convex and generally NP-hard. When the grid is large and thus involves many control variables, traditional model-based methods may obtain suboptimal solutions, which will consume more time or even be impossible to solve. Additionally, this is a multi-stage planning problem, where the decisions of each type of controller are not made at the same stage. To overcome these difficulties, a model-free

method based on deep reinforcement learning is introduced to solve the problem, which is detailed in Section 3.

### 3. Deep Reinforcement Learning Solution

In this section, the voltage control problem is first formulated as an MDP, and then a model-free solution based on deep reinforcement learning is put forward, in which the control variables of different controllers are assigned to different agents. The solution of discrete variables is based on a DQN agent while the solution of continuous variables is based on a DDPG agent.

#### 3.1. Markov Decision Process

In order to solve the voltage control problem with DRL algorithms, the optimal configurations of different controllers have to be modeled as an MDP. An MDP is defined by the tuple  $(S, A, P, R, \gamma)$ , and it is used to describe the interaction process between the agents (i.e., different controllers) and the environment (i.e., the power flow of distribution systems). In this paper, for each agent, the state space  $S$  is continuous while the action space  $A$  is either discrete or continuous.  $P$ , usually unknown, is the state transition probability indicating the probability density of the next state  $s_{t+1} \in S$  under the current state  $s_t$  and action  $a_t$ .  $R$  is the reward on each transition, which is denoted as  $r_t = R(s_t, a_t)$ , and  $\gamma \in [0, 1]$  is the discount factor. Then, the goal of the voltage control problem is to solve the MDP—that is, to learn the optimal policy of each agent to maximize the reward, which is associated with the long-term average voltage deviation.

In DRL algorithms, the policy  $\mu$ , expressed as  $\mu(a|s)$ , is a mapping function of the action  $a_t$  taken by the agent in the state  $s_t$ . During the training process, the action-value function, also called the Q-function, represents the expected discounted reward after taking action  $a$  in the state  $s$  with policy  $\mu$ , and can be denoted as  $Q_\mu(s, a) = \mathbb{E}_\mu \left[ \sum_{k=0}^{T_{all}} \gamma^k r_k \mid s_0 = s, a_0 = a \right]$ , where  $T_{all}$  is the episode length. Using the Bellman equation, the Q-function can be further expressed as  $Q_\mu(s_t, a_t) = \mathbb{E}_\mu [r_t + \gamma Q_\mu(s_{t+1}, a_{t+1}) \mid s_t, a_t]$ . Then, solving the optimal policy  $\mu^*$  is equivalent to solving the optimal Q-function, that is,  $Q_{\mu^*} = \max_{\mu} Q_\mu(s, a)$ .

#### 3.2. DQN-Based Agent for Discrete Variables

The configurations of the capacitors are made at the beginning of each interval  $T$ . For the discrete variables of capacitors, a DQN—a value-based DRL method—is introduced to handle the control problem with continuous state space and discrete action space.

The classic DQN method, based on the Q-learning method, uses a deep neural network (DNN) to estimate the continuous Q-function, and the DNN can be indicated as a Q network, that is,  $Q_\mu(s, a; \theta_Q)$ , whose input is the state vector and output is the Q-values for all possible actions. The experience replay buffer  $D$  is used to store experiences  $e_T = (s_T, a_T, r_T, s_{T+1})$ , and a mini batch is applied to store  $M$  randomly sampled experiences. In order to update the parameters of the Q network, a target Q network is employed with the parameters of  $\theta'_Q$ . Then, using stochastic gradient descent (SGD), the parameters of the Q network can be updated based on the mini batch and the loss function, which can be expressed as

$$L(\theta_Q) = \mathbb{E}_{i \in M} \left[ \left( r_i + \gamma \max_{a'} Q_\mu(s_{i+1}, a'; \theta'_Q) - Q_\mu(s_i, a_i; \theta_Q) \right)^2 \right], \quad (17)$$

where the parameters of the target Q network  $\theta'_Q$  are updated by copying the parameters of the Q network  $\theta_Q$  periodically for every  $B$ .

For the capacitor agent, the state consists of the average positive power of each bus during the interval  $T$  and the action during the last interval  $T-1$ , that is,  $s_{cap}(T) = [\bar{P}^T(T), a_{cap}^T(T-1)]^T$ . The action  $a_{cap}(T) = [a_{cap,1}(T), \dots, a_{cap,N_{cap}}(T)]^T$  is defined as the

configurations of capacitors, where  $a_{cap,i}(T) \in \{0, 1\}$  and  $N_{cap}$  is the number of capacitors. When the state  $s_{cap}(T)$  is fed into the input layer and passes through the hidden layers, the output layer generates the Q-values of the particular configurations of all capacitors with a total number of  $2^{N_{cap}}$  neurons. Then, the action having the maximum Q-value is selected for the next interval. To meet the control object, the reward is designed as the negative of the voltage deviation of all buses, which can be expressed as  $r_{cap}(T) = - \sum_{t=1}^{N_t} \sum_{i=1}^N (U_i(T, t) - 1)^2$ .

In order to ensure that the agent can both explore the unknown environment and make use of the knowledge it has already grasped, the  $\epsilon$ -greedy strategy is employed to select action; that is,

$$a_T = \begin{cases} \text{random } A & \beta < \epsilon \\ \underset{a_T \in A}{\operatorname{argmax}} Q_\mu(s_T, a_T; \theta_Q) & \beta \geq \epsilon \end{cases} \quad (18)$$

where  $\epsilon \in [0, 1]$  and is a constant, and  $\beta \in [0, 1]$  and is randomly generated by computer. When  $\beta < \epsilon$ , the agent randomly selects an action in the action space; otherwise, the agent selects the action that has the maximum Q-value in the current state.

The capacitors' configuration based on the DQN is depicted in Figure 1. During the training period, the agent selects actions based on (18), while during the execution process, based on the current state, the agent selects the action that has the maximum Q-value.

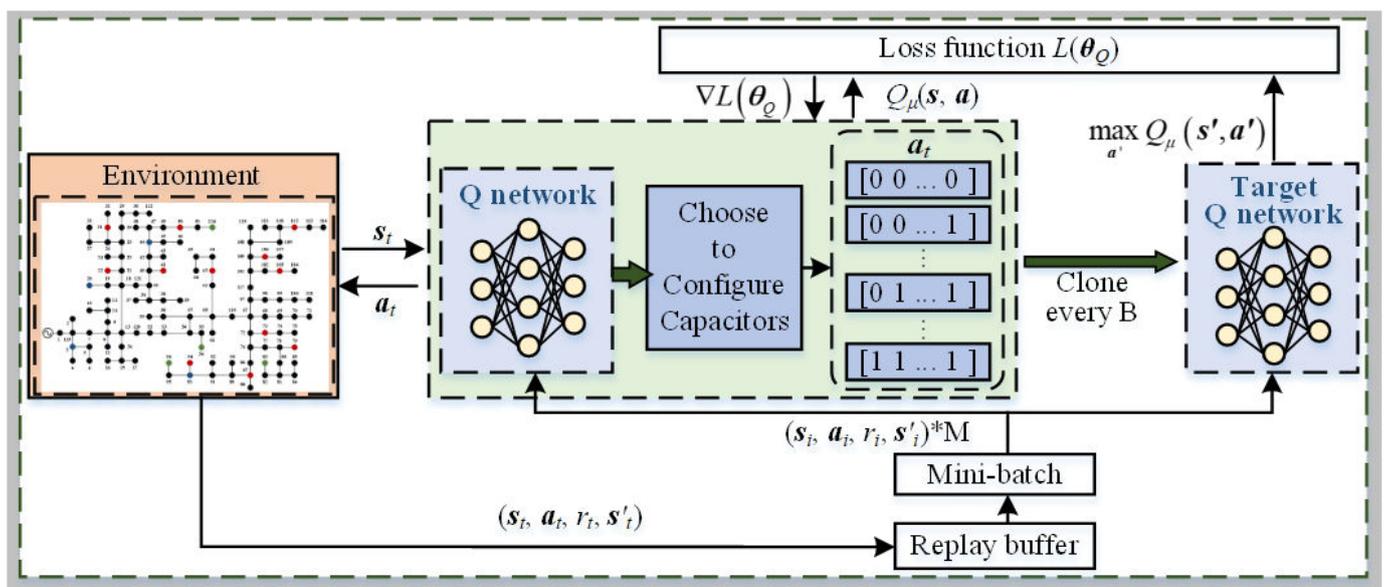


Figure 1. The control of capacitors based on the DQN agent.

### 3.3. DDPG-Based Agent for Continuous Variables

Based on the configuration of capacitors in the intervals, the outputs of PV inverters and batteries are adjusted at the beginning of each slot  $t$ . For the continuous control variables of PV inverters and batteries, DDPG is applied to deal with the control problem with continuous state space and continuous action space.

The DDPG method not only employs a DNN to simulate the Q-function, but also uses a DNN to estimate the policy function. It adopts a typical actor–critic framework, which realizes the policy action and action evaluation by designing the actor network  $\mu(s; \theta_\mu)$  and critic network  $Q_\mu(s, a; \theta_Q)$ , respectively. Like the DQN, the target actor network  $\mu'(s, \theta_{\mu'})$  and target critic network  $Q'_{\mu'}(s, a; \theta_{Q'})$  are also applied.

During the training process, the continuous action is decided based on the following function:

$$\mathbf{a}_t = \mu(\mathbf{s}_t; \boldsymbol{\theta}_\mu) + \zeta_t \quad (19)$$

where  $\zeta_t$  is the noise used to randomly search actions in the action space. Experience replay buffer and mini batch are also employed. Then, the critic network can be updated by minimizing the loss function as in (17); that is,

$$L(\boldsymbol{\theta}_Q) = \mathbb{E}_{i \in M} [y_i - Q_\mu(\mathbf{s}_i, \mathbf{a}_i; \boldsymbol{\theta}_Q)]^2 \quad (20)$$

$$y_i = r_i + \gamma Q'_\mu(\mathbf{s}_{i+1}, \mu'(\mathbf{s}_{i+1}; \boldsymbol{\theta}_{\mu'}); \boldsymbol{\theta}_{Q'}) \quad (21)$$

The actor network can be updated using the policy gradient, which can be expressed as

$$\nabla_{\boldsymbol{\theta}_\mu} \mu = \mathbb{E}_{i \in M} [\nabla_a Q(\mathbf{s}_i, \mathbf{a}_i; \boldsymbol{\theta}_Q) \cdot \nabla_{\boldsymbol{\theta}_\mu} \mu(\mathbf{s}_i; \boldsymbol{\theta}_\mu)] \quad (22)$$

Then, the target networks are soft-updated as follows:

$$\boldsymbol{\theta}'_Q = \lambda \boldsymbol{\theta}_Q + (1 - \lambda) \boldsymbol{\theta}'_Q \quad (23)$$

$$\boldsymbol{\theta}'_\mu = \lambda \boldsymbol{\theta}_\mu + (1 - \lambda) \boldsymbol{\theta}'_{\mu'} \quad (24)$$

where the parameter  $\lambda \ll 1$ .

For the PV inverters and batteries agent, the state consists of the voltage amplitude of all buses and the state of charge of batteries at time  $t$ , that is,  $\mathbf{s}_{PVbatt}(t) = [\mathbf{U}^T(t), \mathbf{SOC}^T(t)]^T$ . The action includes the action variables of PV inverters and batteries, that is,  $\mathbf{a}_{PVbatt}(t) = [\mathbf{a}_{PV}^T(t), \mathbf{a}_{batt}^T(t)]^T$ ,  $\mathbf{a}_{PV}(t) = [a_{PV,1}(t), \dots, a_{PV,N_{PV}}(t)]^T$ , and  $\mathbf{a}_{batt}(t) = [a_{batt,1}(t), \dots, a_{batt,N_{batt}}(t)]^T$ , where  $N_{PV}$  and  $N_{batt}$  are the number of PV inverters and batteries, respectively, and  $a_{PV,i}(t) \in [-1, 1]$  and  $a_{batt,i}(t) \in [-1, 1]$ . The reward is designed similarly to that of the capacitors agent, which can be denoted as  $r_{PVbatt}(t) = -\sum_{i=1}^N (U_i(t) - 1)^2$ .

Furthermore, the action implementation method of forced constraint output is adopted in order to take the capacity boundaries of batteries into account—that is, when  $SOC_i(t+1) = SOC_i(t) + a_{PV,i}(t) \cdot P_{batt,i}^{max}$  is outside of the upper and lower boundaries, only the amount of chargeable (dischargeable) power  $SOC_{i,max} - SOC_i(t)$  ( $SOC_i(t) - SOC_{i,min}$ ) is charged (discharged).

The control of PV inverters and batteries based on DDPG is demonstrated in Figure 2. During the training period, the state  $\mathbf{s}_{PVbatt}(t)$  is fed into the actor network and an action  $\mathbf{a}_{PVbatt}(t)$  is generated based on (19). Then, the state and action enter the critic network and the corresponding  $Q$ -value is generated. During the execution period, with well-trained networks, the agent chooses its action based on the state, that is,  $\mathbf{a}_{PVbatt}(t) = \mu(\mathbf{s}_{PVbatt}(t); \boldsymbol{\theta}_\mu)$ .

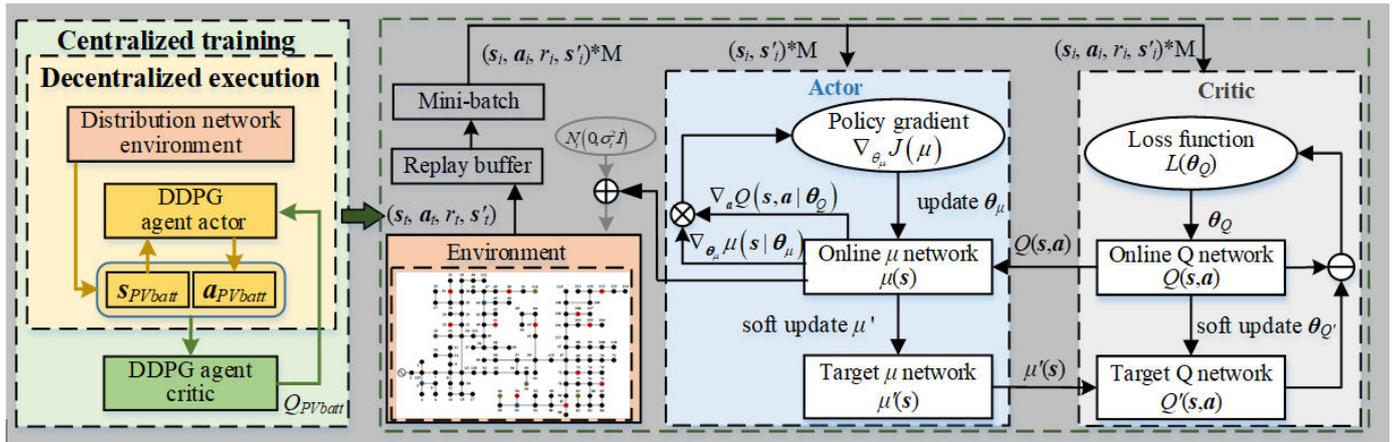


Figure 2. The control of PV inverters and batteries based on DDPG agent.

### 3.4. Algorithm and Computation Process

The two-timescale voltage control for distribution systems based on DRL is demonstrated in Algorithm 1.

**Algorithm 1** Two-timescale voltage control scheme for distribution systems based on DRL.

- 1: For the capacitors agent based on DQN, initialize the parameters  $\theta_{Q,DQN}$  randomly; initialize the target Q network with  $\theta'_{Q,DQN} = \theta_{Q,DQN}$ ; initialize replay buffer  $D_{DQN}$
- 2: For the PV inverters and batteries agent based on DDPG, initialize the actor and critic networks with  $\theta_{\mu,DDPG}$  and  $\theta_{Q,DDPG}$ ; initialize target actor and critic network with  $\theta'_{\mu,DDPG} = \theta_{\mu,DDPG}$ ,  $\theta'_{Q,DDPG} = \theta_{Q,DDPG}$ ; initialize replay buffer  $D_{DDPG}$
- 3: Initialize capacitors agent state  $s_{DQN1}$
- 4: **for**  $T = 1$  to  $N_T$  **do**
- 5:   Get the action  $a_{capT}$  from (18)
- 6:   Execute  $a_{capT}$  in the power flow environment and get the initiation state of DDPG agent:  $s_{PVbattT1}$
- 7:   **for**  $t = 1$  to  $N_t$  **do**
- 8:     Get the action  $a_{PVbattTt}$  from (19); execute  $a_{PVbattTt}$  in the power flow environment and get the reward  $r_{PVbattTt}$  and new state  $s'_{PVbattTt}$
- 9:     Store the experience in the replay buffer  $D_{DDPG}$
- 10:     Sample a random mini batch from  $D_{DDPG}$  and update the actor and critic networks using (20), (21) and (22).
- 11:     Soft update the target actor and critic networks using (23) and (24)
- 12:   **end for**
- 13:   Get the reward of capacitors  $r_{capT}$ ; get the new state  $s'_{capT}$
- 14:   Store the experience in the replay buffer  $D_{DQN}$
- 15:   Sample a mini batch randomly from  $D_{DQN}$  and update the Q network for DQN
- 16:   **if**  $\text{mod}(T,B) = 0$  **then**
- 17:     Update the target Q network for DQN:  $\theta'_{Q,DQN} = \theta_{Q,DQN}$
- 18:   **end if**
- 19: **end for**

## 4. Numerical Study

In this section, the implementation details of the proposed two-timescale voltage control scheme based on DRL are described.

### 4.1. Simulation Setup

A modified IEEE 123-bus distribution test system was applied to carry out the numerical tests. Based on the original 123-bus multi-phase unbalanced network [21], the system was changed into a balanced system and the numbering of each node was reorganized as shown in Figure 3. Twelve PV units with smart inverters were installed at 12 buses, and their capacities and locations are listed in Table 1. Four capacitors were installed in the grid at buses 20, 59, 66, and 114, each with a capacity of 40 kvar. Four energy storage batteries

were installed at buses 56, 83, 96, and 116, each with a maximum capacity of 600 kWh and rated charge/discharge power of 100 kW. The load power was modified from the real data of an area in Jiangsu province, China (i.e., on each bus, the peak load value was set to the sum of the loads on three phases in the original 123-bus distribution network, and the load curve after standardization was the same as that after the standardization of the real load of an area in Jiangsu province, China). Thus, the load value of each bus was equal to the load curve multiplied by its peak load. All parameters in this distribution system were converted to a consistent base, where the base voltage was 4.16 kV and the base power was 100 MVA.

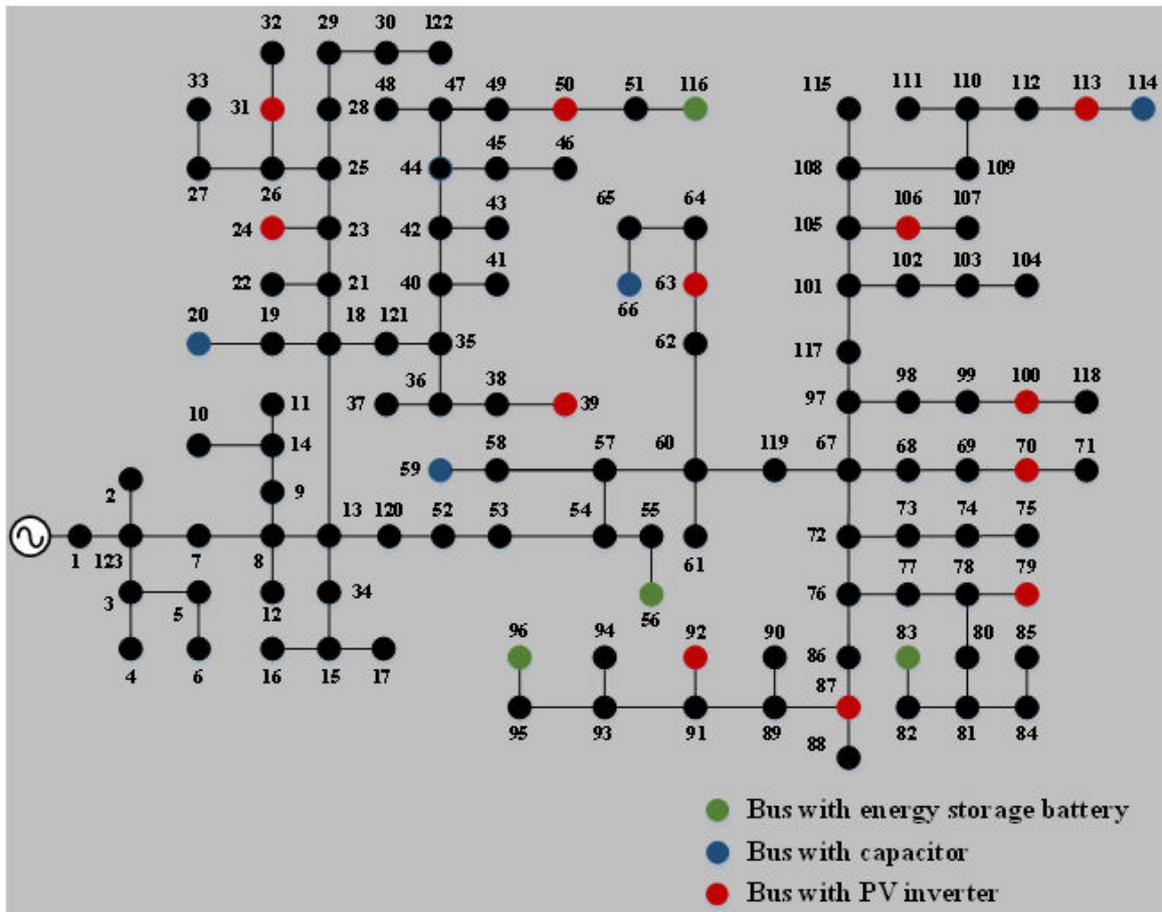


Figure 3. The modified IEEE 123-bus system topology.

Table 1. The capacities and locations of 12 installed PV units.

PV Location	PV Capacity	PV Location	PV Capacity	PV Location	PV Capacity
24	300 kW	63	300 kW	92	400 kW
31	300 kW	70	400 kW	100	400 kW
39	200 kW	79	400 kW	106	300 kW
50	400 kW	87	200 kW	113	200 kW

In the theory of DRL, the parameters that define the architecture of the NNs are of great importance, and the selection of architecture depends on the actual application scenario. For example, convolutional neural networks (CNNs) are often used to deal with complex problems in the image domain, whereas recurrent neural networks (RNNs) are often used to process sequence data. For the voltage control problem raised in this paper, a fully connected NN was sufficient for the task at hand. Based on [22], the number of hidden layers was chosen to be 2 and the number of neurons in each hidden layer was

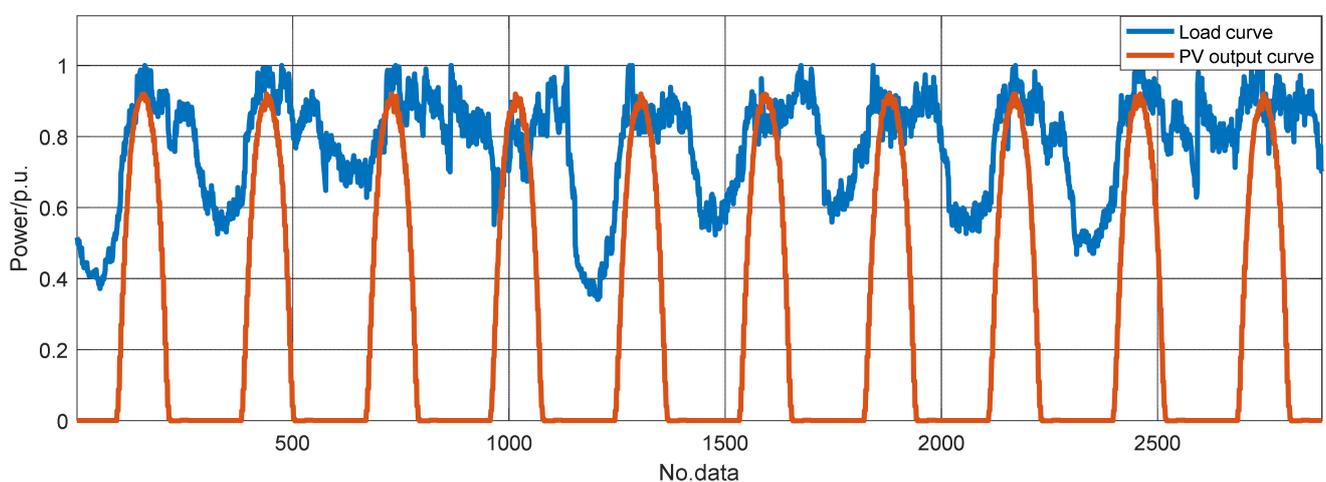
first selected according to that in the upper layer and the lower layer. Then, the model was trained, the output was checked to see if there was overfitting, and the parameters were adjusted until a satisfactory output was obtained. According to the above system setting, the four capacitors generated  $2^4 = 16$  combinations of discrete actions, and the PV inverters and batteries produced 16 continuous actions. For the DQN agent, the Q network consisted of three fully connected layers: one input layer, two hidden layers with 95 and 22 neurons, respectively, and one output layer with 16 neurons. The sigmoid function was used at the end of the output layer to keep the Q-value within  $[0, 1]$ . For the DDPG agent, the actor and critic networks were also composed of three fully connected layers, with hidden layers of 90, 30 units, and 46, 14 units, respectively. The output layer of the actor network consisted of 16 neurons and the output layer of the critic network had 1 neuron. The tanh function was applied at the end of the actor network to keep the action variables within  $[-1, 1]$ . All the hidden layers used rectified linear unit (ReLU) as the activation function. The detailed settings of other hyper-parameters are declared in Table 2. Optimal power flow was employed as the environment for these DRL agents. The proposed algorithm was run in Python using the Pytorch framework, and the training process was executed on CPU.

**Table 2.** Settings of other DRL parameters.

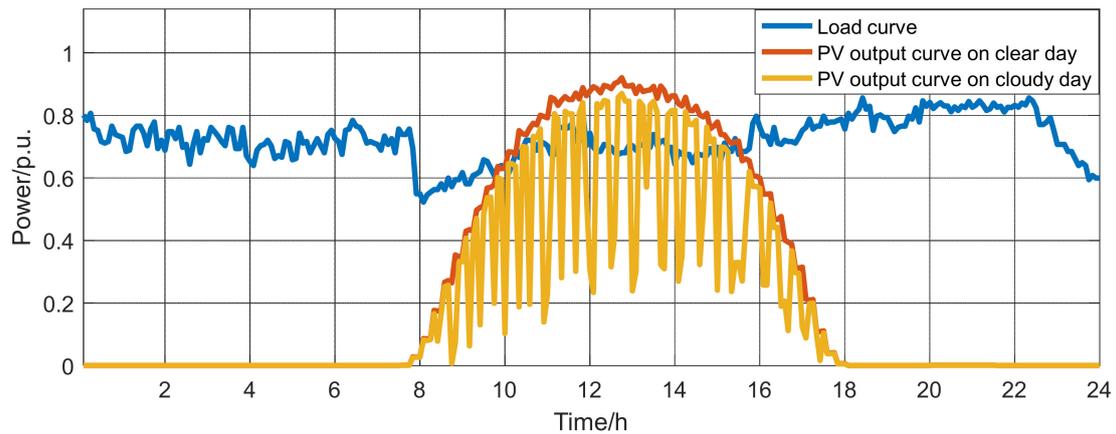
Item	Value
Replay buffer size for DQN agent	50
Mini-batch size for DQN agent	8
Target Q network updating cycle	10
Learning rate for DQN network	0.01
Replay buffer size for DDPG agent	1000
Mini-batch size for DDPG agent	64
Learning rate for critic network	0.001
Learning rate for actor network	0.001
Discount factor	0.99
Soft-updating parameter $\lambda$	0.005

#### 4.2. Case Study

In this subsection, we evaluate the performance of the proposed DRL scheme using the modified IEEE 123-bus system. A total of 2880 data points, comprising load data and PV outputs, were used as training data, as demonstrated in Figure 4. Meanwhile, another 288 data points were used as test data, as depicted in Figure 5.

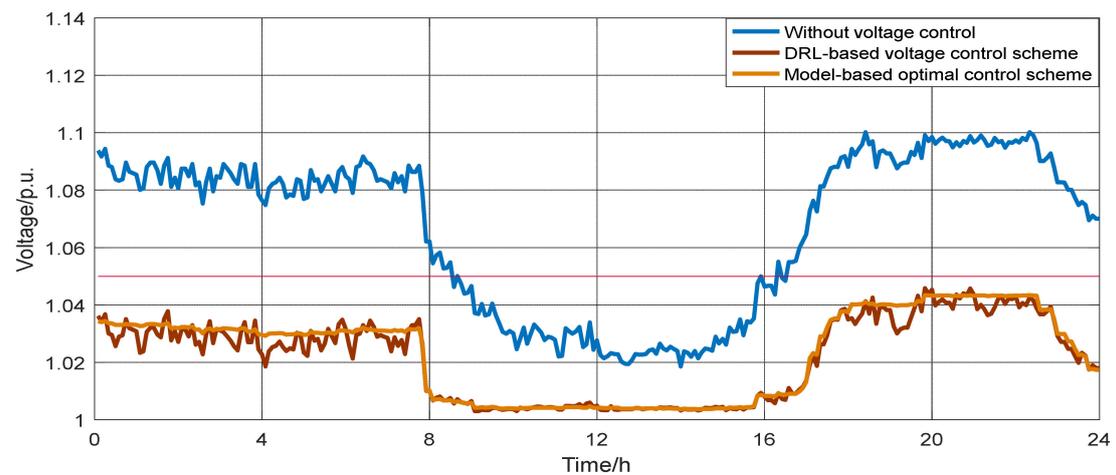


**Figure 4.** The load data and PV output data in the training set.

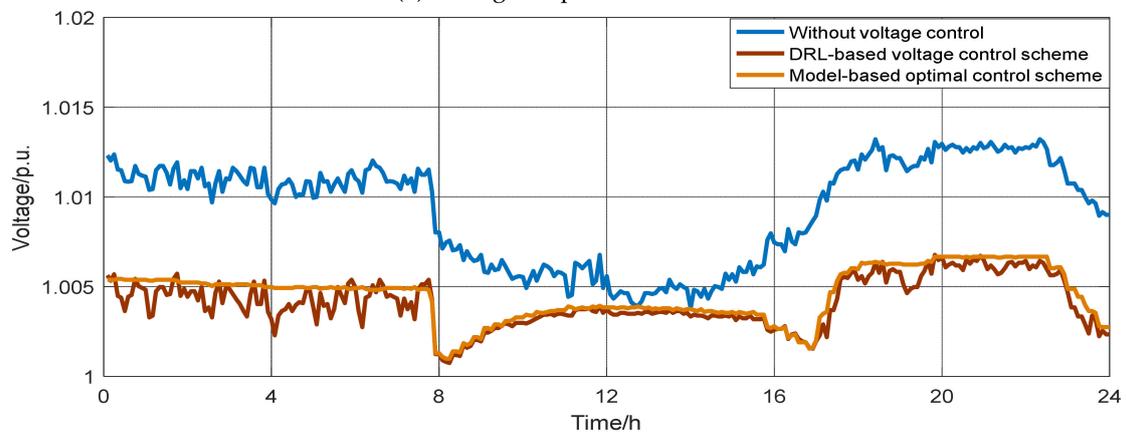


**Figure 5.** The load data and PV output data in the test set.

First, based on the optimal power flow, the voltage distribution of this grid without any voltage control was analyzed. In this paper, the interval  $T$  was defined as 30 min and the slot  $t$  was assumed to be 5 min. The PV outputs were based on the clear day in Figure 5. The buses experiencing voltage issues were bus 1, bus 2, bus 7, and bus 123, which violated the maximum voltage limit of 1.05. Take for example the voltage amplitudes at bus 1 and bus 24 (with PV unit installation), which are depicted by the blue curve in Figure 6.



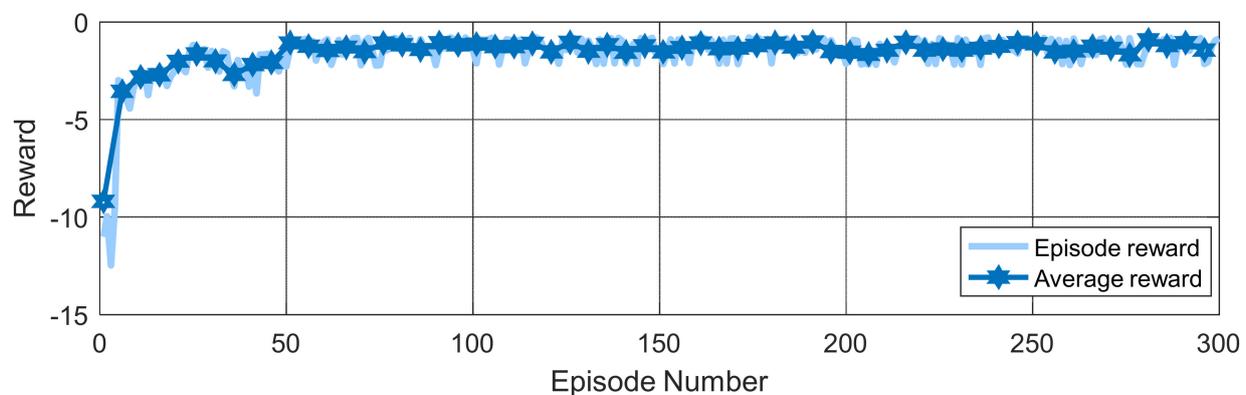
**(a)** Voltage amplitude at bus 1



**(b)** Voltage amplitude at bus 24

**Figure 6.** Voltage profiles under different methods on a clear day. (a) Voltage amplitude at bus 1, (b) Voltage amplitude at bus 24.

Then, the learning performance of the proposed DRL method was investigated. Following the procedure shown in Algorithm 1, the DQN and DDPG agents were trained. During the training process, the daily PV generation and load consumption combinations were randomly chosen from the training set, as demonstrated in Figure 4, to represent different grid operation conditions. Training was performed for 300 episodes, and each episode finished after training 288 samples from one day. Figure 7 displays the episode reward values and average rewards in the training period, where the episode reward value is the sum of all the rewards obtained during a given episode and the average reward value is the average of every four episode reward values. It can be observed that in the early stages the reward value was very low because of the limited learning experiences. As the training process continued, the agents gradually evolved and the reward value increased. Later, the reward curves fluctuated due to the random control attempts of both the DQN and DDPG agents to determine the correct policy actions. After about 60 episodes, the reward curve flattened out gradually, indicating the DRL agents' ability to realize voltage control.



**Figure 7.** The rewards of DRL agents during the training process.

During the test period, the trained DRL agents were employed to control the capacitors, batteries, and PV inverters according to the test data in Figure 5, and the PV outputs were based on the clear day. As demonstrated in Figure 6, compared with the case without voltage control, these trained DRL agents demonstrated an effective performance for voltage control and all bus amplitudes were maintained within the safety limit, especially the buses having voltage issues. Thus, we can conclude that the proposed algorithm enables the controllers to explore the relationship between their configurations and the inherent uncertainty and variability in the PV outputs and load power, and to take corresponding policies when faced with new operating conditions.

#### 4.3. Comparison with the Model-Based Optimal Control Scheme

In order to compare the performance of the proposed DRL method, a model-based optimal control scheme called a two-stage optimal control scheme was applied. The model-based optimal control scheme aims to minimize the daily voltage deviation, and was assumed to have full knowledge of the model and parameters of the distribution network. In this method, the configurations of capacitors, batteries, and PV inverters were decided at the beginning of each interval and the outputs of batteries and PV inverters were further adjusted based on the capacitors configuration at the beginning of each slot.

As demonstrated in Figure 6, in most cases the control effect of the DRL-based method was similar to or even better than that of model-based method, since the model-based method considers the optimal control throughout a day, while the DRL-based method can realize real-time control according to the current state of the power grid.

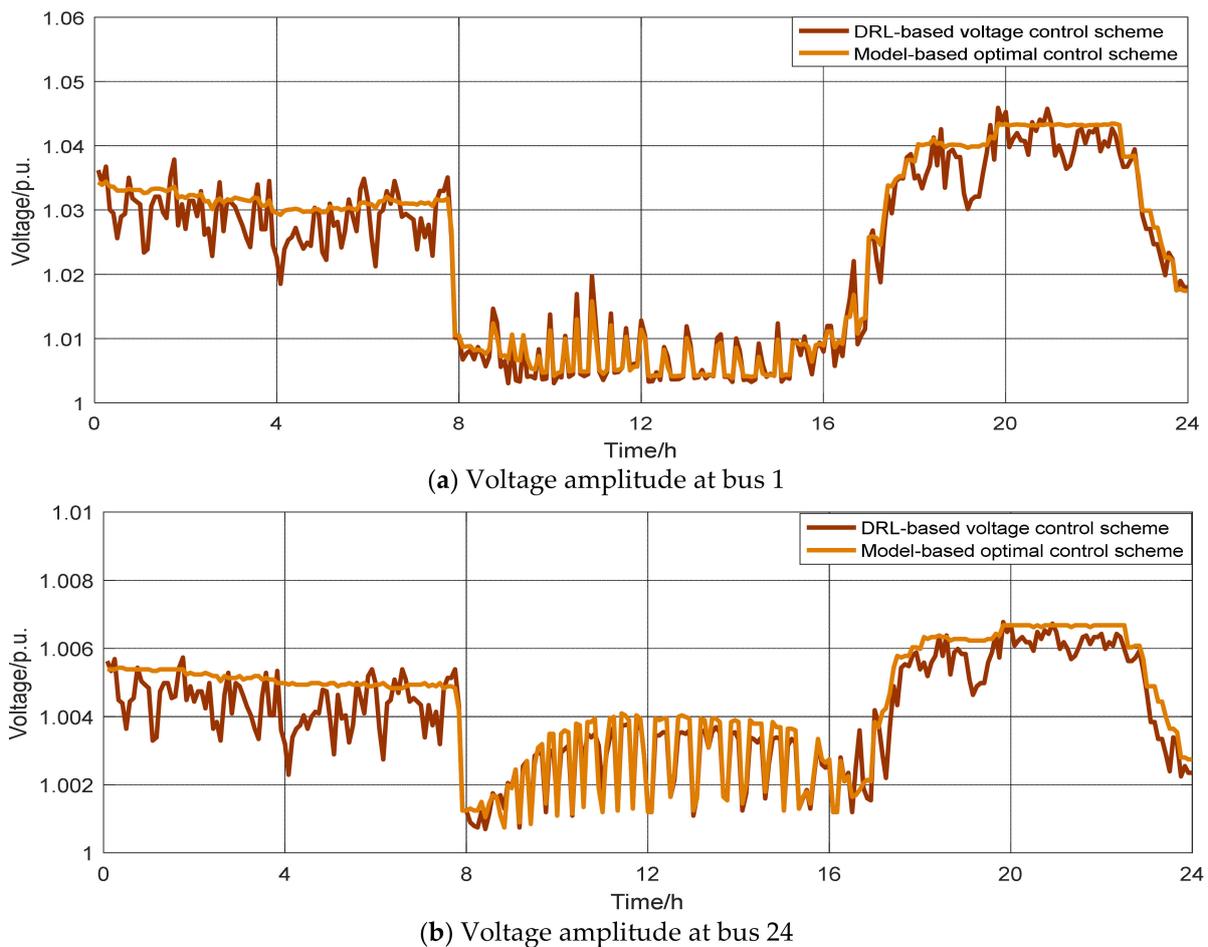
The execution time of the model-based control method and our proposed DRL-based method targeting all of the day's 288 samples is demonstrated in Table 3. It can be seen that the proposed DRL-based method took only 0.1964 s, much less than the 1107.7629 s of

the model-based method. Therefore, the proposed algorithm shows high computational efficiency and has competitive potential in online application.

**Table 3.** Execution time of model-based control method and DRL-based method.

Method	Time (s)
Model-based control method	1107.7629
Our proposed DRL-based control method	0.1964

In order to evaluate the dynamic response performance of the proposed DRL-based controller under time-varying PV outputs, the case of a cloudy day was studied. The outputs of the PV units were based on the cloudy day in Figure 5, which show a great deal of fluctuation. The results of the model-based optimal control scheme and the proposed DRL-based voltage control scheme are depicted in Figure 8. It can be seen that the proposed DRL-based controller could respond quickly to the PV fluctuations, which is very important in order to realize the demand of real-time control. Additionally, in the model-based scheme, the controller needs prior knowledge of the PV outputs over a period of time, which is often inaccurate in the case of PV fluctuations. In the proposed DRL-based scheme, the controller adjusts its action based on the current state of the power grid, and is more reliable.



**Figure 8.** Voltage profiles under different methods on a cloudy day. (a) Voltage amplitude at bus 1, (b) Voltage amplitude at bus 24.

In order to evaluate the control performance of the proposed DRL-based voltage controller under extreme weather conditions, a case study was carried out for a scenario

where there were no PV outputs at all. The results of the model-based optimal control scheme and the proposed DRL-based voltage control scheme are demonstrated in Figure 9. The voltage amplitudes of all buses were controlled below 1.05 and it can be seen that the proposed DRL-based controller still had better control performance without PV outputs. From the results of Figures 7–9, it can be observed that the trained DRL-based controller worked very well in different scenarios, and could adapt to similar but slightly different data, which verifies the generalization ability of the proposed algorithm.

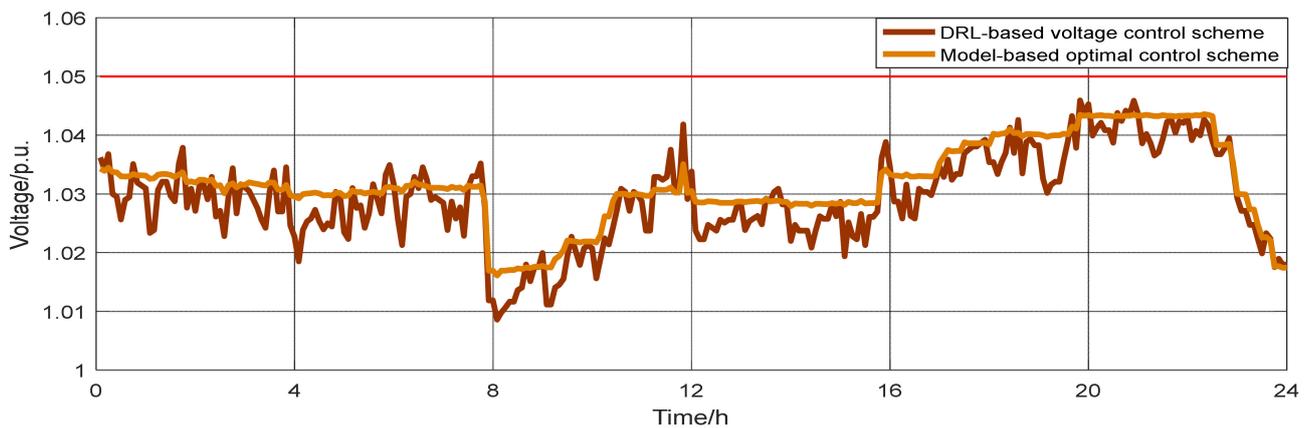


Figure 9. Voltage amplitude at bus 1 under different methods with no PV outputs.

## 5. Conclusions

In this paper, a two-timescale voltage control scheme based on a DRL method is proposed to control multiple types of equipment, including capacitors, energy storage batteries, and PV inverters, for optimal voltage control in the distribution network. Control variables are assigned to different agents according to their properties, which share the same environment and are trained simultaneously to cooperate with each other. Specifically, the discrete variables are solved using a DQN agent and the continuous variables are solved using a DDPG agent. A specially designed reward is applied to achieve the goal of minimizing long-term average voltage deviation. Case studies showed that the proposed algorithm had similar or even better performance than the model-based optimal control scheme, and had high computational efficiency, enabling the realization of real-time control. Additionally, the proposed DRL-based controller could adjust its action based on the current state of the power grid. It had better dynamic response performance and could enact a quick response to PV fluctuations. Future work will focus on designing the reward function to achieve more control objectives and take various operating constraints into consideration.

**Author Contributions:** Conceptualization, J.Z.; formal analysis, Y.L.; funding acquisition, T.W.; investigation, Y.L.; methodology, J.Z.; project administration, C.R.; resources, Z.Z.; software, Y.L.; supervision, Z.W. and S.Z.; validation, J.Z. and Z.W.; writing—original draft, Y.L.; writing—review and editing, Z.W. and S.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the science and technology project of the State Grid Corporation of China, Grant 5204JY20000B.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data sharing not applicable. No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

## References

1. Zhang, L.; Chen, Y.; Shen, C.; Tang, W.; Liang, J. Coordinated voltage regulation of hybrid AC/DC medium voltage distribution networks. *J. Mod. Power Syst. Clean Energy* **2018**, *6*, 463–472. [[CrossRef](#)]
2. Hu, X.; Liu, Z.-W.; Wen, G.; Yu, X.; Liu, C. Voltage Control for Distribution Networks via Coordinated Regulation of Active and Reactive Power of DGs. *IEEE Trans. Smart Grid* **2020**, *11*, 4017–4031. [[CrossRef](#)]
3. Tang, Z.; Hill, D.J.; Liu, T. Fast Distributed Reactive Power Control for Voltage Regulation in Distribution Networks. *IEEE Trans. Power Syst.* **2019**, *34*, 802–805. [[CrossRef](#)]
4. Xie, Y.; Liu, L.; Wu, Q.; Zhou, Q. Robust model predictive control based voltage regulation method for a distribution system with renewable energy sources and energy storage systems. *Int. J. Electr. Power Energy Syst.* **2020**, *118*, 105749. [[CrossRef](#)]
5. Chen, S.; Hu, W.; Chen, Z. Comprehensive Cost Minimization in Distribution Networks Using Segmented-Time Feeder Reconfiguration and Reactive Power Control of Distributed Generators. *IEEE Trans. Power Syst.* **2016**, *31*, 983–993. [[CrossRef](#)]
6. Sun, K.M.; Chen, Q.; Zhao, P. Genetic algorithm with mesh check for distribution network topology reconfiguration. *Autom. Electr. Power Syst.* **2018**, *42*, 64–71. (In Chinese)
7. Zhao, H.; Zhao, J.; Qiu, J.; Liang, G.; Dong, Z.Y. Cooperative Wind Farm Control with Deep Reinforcement Learning and Knowledge-Assisted Learning. *IEEE Trans. Ind. Inform.* **2020**, *16*, 6912–6921. [[CrossRef](#)]
8. Ye, Y.; Qiu, D.; Li, J.; Strbac, G. Multi-Period and Multi-Spatial Equilibrium Analysis in Imperfect Electricity Markets: A Novel Multi-Agent Deep Reinforcement Learning Approach. *IEEE Access* **2019**, *7*, 130515–130529. [[CrossRef](#)]
9. Li, H.Z.; Wang, L.; Lin, D.; Zhang, X.Y. A nash game model of multi-agent participation in renewable energy consumption and the solving method via transfer reinforcement learning. *Proc. CSEE* **2019**, *39*, 4135–4149. (In Chinese)
10. Cao, J.; Harrold, D.; Fan, Z.; Morstyn, T.; Healey, D.; Li, K. Deep Reinforcement Learning-Based Energy Storage Arbitrage with Accurate Lithium-Ion Battery Degradation Model. *IEEE Trans. Smart Grid* **2020**, *11*, 4513–4521. [[CrossRef](#)]
11. Li, H.; Wan, Z.; He, H. Constrained EV Charging Scheduling Based on Safe Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2019**, *11*, 2427–2439. [[CrossRef](#)]
12. Hojabri, M.; Dersch, U.; Papaemmanouil, A.; Bosshart, P. A Comprehensive Survey on Phasor Measurement Unit Applications in Distribution Systems. *Energies* **2019**, *12*, 4552. [[CrossRef](#)]
13. Shang, X.Y.; Li, M.S.; Ji, T.Y.; Zhang, L.L.; Wu, Q.H. Discrete reactive power optimization considering safety margin by dimensional Q-learning. In Proceedings of the 2015 IEEE Innovative Smart Grid Technologies—Asia (ISGT ASIA), Bangkok, Thailand, 3–6 November 2015; pp. 1–5.
14. Yang, Q.; Wang, G.; Sadeghi, A.; Giannakis, G.B.; Sun, J. Two-Timescale Voltage Control in Distribution Grids Using Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2020**, *11*, 2313–2323. [[CrossRef](#)]
15. Lukianychin, O.; Bogodorova, T. Voltage Control-Based Ancillary Service Using Deep Reinforcement Learning. *Energies* **2021**, *14*, 2274. [[CrossRef](#)]
16. Zhang, Y.; Wang, X.; Wang, J.; Zhang, Y. Deep Reinforcement Learning Based Volt-VAR Optimization in Smart Distribution Systems. *IEEE Trans. Smart Grid* **2020**, *12*, 361–371. [[CrossRef](#)]
17. Duan, J.; Shi, D.; Diao, R.; Li, H.; Wang, Z.; Zhang, B.; Bian, D.; Yi, Z. Deep-Reinforcement-Learning-Based Autonomous Voltage Control for Power Grid Operations. *IEEE Trans. Power Syst.* **2020**, *35*, 814–817. [[CrossRef](#)]
18. Sun, X.; Qiu, J. Two-Stage Volt/Var Control in Active Distribution Networks with Multi-Agent Deep Reinforcement Learning Method. *IEEE Trans. Smart Grid* **2021**, *1*. [[CrossRef](#)]
19. Beyer, K.; Beckmann, R.; Geißendörfer, S.; von Maydell, K.; Agert, C. Adaptive Online-Learning Volt-Var Control for Smart Inverters Using Deep Reinforcement Learning. *Energies* **2021**, *14*, 1991. [[CrossRef](#)]
20. Al-Saffar, M.; Musilek, P. Reinforcement Learning-Based Distributed BESS Management for Mitigating Overvoltage Issues in Systems with High PV Penetration. *IEEE Trans. Smart Grid* **2020**, *11*, 2980–2994. [[CrossRef](#)]
21. Kersting, W.H. Radial distribution test feeders. In Proceedings of the IEEE Power Engineering Society Winter Meeting, Columbus, OH, USA, 28 January–1 February 2001; Volume 2, pp. 908–912. [[CrossRef](#)]
22. Heaton, J. Heaton Research the Number of Hidden Layers. 2017. Available online: <https://www.heatonresearch.com/2017/06/01/hidden-layers.html> (accessed on 9 June 2021).