Intelligent Voltage Control Method in Active Distribution Networks Based on Averaged Weighted Double Deep Q-network Algorithm

Yangyang Wang, *Student Member*, *IEEE*, Meiqin Mao, *Senior Member*, *IEEE*, Liuchen Chang, *Senior Member*, *IEEE*, and Nikos D. Hatziargyriou, *Fellow*, *IEEE*

***Abstract*—High penetration of distributed renewable energy sources and electric vehicles (EVs) makes future active distribution network (ADN) highly variable. These characteristics put great challenges to traditional voltage control methods. Voltage control based on the deep Q-network (DQN) algorithm offers a potential solution to this problem because it possesses human-level control performance. However, the traditional DQN methods may produce overestimation of action reward values, resulting in degradation of obtained solutions. In this paper, an intelligent voltage control method based on averaged weighted double deep Q-network (AWDDQN) algorithm is proposed to overcome the shortcomings of overestimation of action reward values in DQN algorithm and underestimation of action reward values in double deep Q-network (DDQN) algorithm. Using the proposed method, the voltage control objective is incorporated into the designed action reward values and normalized to form a Markov decision process (MDP) model which is solved by the AWDDQN algorithm. The designed AWDDQN-based intelligent voltage control agent is trained offline and used as online intelligent dynamic voltage regulator for the ADN. The proposed voltage control method is validated using the IEEE 33-bus and 123-bus systems containing renewable energy sources and EVs, and compared with the DQN and DDQN algorithms based methods, and traditional mixed-integer nonlinear program based methods. The simulation results show that the proposed method has better convergence and less voltage volatility than the other ones.**

***Index Terms*—Averaged weighted double deep Q-network (AWDDQN), deep Q learning, active distribution network (ADN), voltage control, electrical vehicle (EV).**

I. Introduction

With the large-scale integration of distributed renewable energy sources (RESs) such as photovoltaic (PV) generators and wind turbines (WTs) and the massive application of electric vehicles (EVs), the power mismatch between the energy production and energy consumption in active distribution networks (ADNs) becomes highly variable. The resulting dynamic changes in power flows and voltages increase the risk of violating the voltage limits at nodes with RESs and EVs of the ADN [1], [2]. These new features of highly variable operational environments make the reactive power and voltage control methods of the traditional distribution network, such as transformer tap-changings [3] and static reactive power compensation [4], highly inefficient. At the same time, with the increasing penetration of RESs, the traditional regulation resources are becoming scarce, requiring the full use of any adjustable resource of the system for voltage control [5]. One such resource is provided by the participation of EVs as a bidirectional and adjustable resource in ADN voltage control using vehicle to grid (V2G) technologies. Through the coordinated control of EVs and other power resources, the adjustable resources can be better utilized, thus reducing grid investment and operation costs [6].

Existing voltage control methods in ADN with EVs usually provide voltage control by regulating EV charging and discharging actions independently or in coordination with other control resources. For example, EV charging power control for the voltage control in ADN is based on model prediction of EV behaviors and convex optimization methods [7]. The work in [8] achieves cost minimization and voltage support at ADN nodes by regulating bidirectional reactive power between the EV and the grid using heuristic algorithms.

However, the randomness of EV state of charge (SOC) and the temporal and spatial uncertainties of the access of EVs to the grid make the schedulable capacity of EVs stochastic and time-varying. It turns the optimal voltage control model using EVs to a stochastic, mixed-integer, and nonlinear programming problem, whose solutions are time-consuming or even infeasible [9].

To solve the above problem, distributed and hierarchical

Manuscript received: March 16, 2022; revised: June 29, 2022; accepted: August 26, 2022. Date of CrossCheck: August 26, 2022. Date of online publication: October 27, 2022.

This work was supported in part by the Anhui Province Natural Science Foundation (No. 2108085UD02), the National Natural Science Foundation of China (No. 51577047), and 111 Project (No. BP0719039).

Y. Wang and M. Mao (corresponding author) are with the Research Center for Photovoltaic System Engineering, Ministry of Education, School of Electrical Engineering and Automation, Hefei University of Technology, Hefei 230009, China (e-mail: wangyywork@126.com; mmqmail@163.com).

L. Chang is with the University of New Brunswick, Fredericton, NB E3B 5A3, Canada (e-mail: lchang@unb.ca).

N. D. Hatziargyriou is with the National Technical University of Athens, Athens 15780, Greece (e-mail: nh@power.ece.ntua.gr).

DOI: 10.35833/MPCE.2022.000146

voltage control methods are proposed in the literature. In particular, hierarchical control of EVs by using EV aggregators (EVAs) is proposed in [9]. In this structure, the optimization variables are changed from a large number of EVs into a few EVAs, which greatly reduces the communicational and computational burden. An EVA may manage one or more charging stations, or a group of EV fleets [10]. For example, the voltage optimization control problem is decomposed into multiple cone optimization problems that are solved separately by using a three-level voltage control strategy and multi-time-scale prediction of EV loads [11]. Alternatively, a Monte Carlo method is used to simulate the EV behaviors and a heuristic algorithm is used to solve the two-level voltage optimization problem [12]. In [13], a discounted stochastic multiplayer game method is used to control the power of EVA resulting from the electricity market auction mechanism to achieve voltage optimization through a hierarchical structure. In [14], voltage and reactive power optimization of EVs is realized using a distributed model predictive control method and optimization results are allocated through a peer-to-peer method.

The model-based voltage control methods mentioned above, however, are highly dependent on prediction results, specific equipment, and optimization models that are difficult to solve accurately. Thus, this class of methods faces difficulties in coping with scenarios of time-varying renewable energy generations, loads, and adjustable capacity of EVs [15].

To address the problems of model-based control methods, model-free control method using reinforcement learning (RL) algorithm has received increasing attention in power system applications, because it does not require a tedious and complicated modeling process, reduces model inaccuracies caused by experience and bias, does not rely on prediction, and has better portability and adaptability. One of the most representative and widely used RL algorithms is the Q-learning. For example, the voltage control methods using Q-learning algorithm have been used for transformer tap and capacitor switching control for the voltage control in ADN [16], [17].

However, traditional Q-learning algorithms based on state-action tables are prone to cause the "curse of dimensionality" in scenarios with many states. So, they cannot efficiently cope with voltage control in an ADN with many nodes and variable states. Deep reinforcement learning (DRL) with human-level control, which combines reinforcement learning with deep learning, such as deep Q-network (DQN) algorithm, can solve the problem with many states by estimating the action reward values of different states [18]. Multiple DQN-based algorithms have been reported in the literature for the voltage control in ADN. For example, the DQN algorithm is used to control the active power from battery energy storage systems and the reactive power from switchable capacitors and inverters to achieve the voltage control in ADN [19], [20]. Multiple DQN algorithms are also combined to achieve multi-time-scale voltage optimization control in ADN [21], and DQN algorithms with deep deterministic policy gradients are used to provide a two-stage EV charging policy to achieve voltage fluctuation suppression of distribution networks [22].

Overestimation or underestimation of actions in DRL can seriously degrade the learning performance. In the DQN algorithm, the same action reward value is used for the selection and evaluation of an action, thus, the problem of overestimation of the action reward value is common [23]. Therefore, a double deep Q-network (DDQN) algorithm is proposed in [23]. This algorithm prevents the overestimation of the action reward value by separating the selection and evaluation of actions when designing the reward target, which has better performance than DQN in Atari 2600 game tests. The voltage control with DDQN algorithm is found to be more effective than conventional proportional control [24]. However, DDQN algorithm also poses the problem of underestimating action reward values, especially in learning environments with stochastic relationships [25]. Therefore, DDQN algorithm may be less effective for power systems dominated by stochastic sources such as RESs, EVs, and loads. Based on this, the averaged weighted double deep Q-network (AWDDQN) [26] algorithm is proposed. This algorithm integrates a dual weighted estimator with weights into DDQN algorithm, thus generating target values from previously learned action estimation values. Thus, the dual weighted estimator reduces the overestimation problem of action reward values of DDQN algorithm. Besides, the impact of the stochastic variations is greatly reduced, overcoming the drawback of DDQN algorithm in underestimating action reward values.

In summary, although the DQN and DDQN algorithms have been used to solve the problem of voltage control in ADN, few of the existing studies discuss the problem of control performance degradation caused by the action estimation bias of DQN and DDQN algorithms. In contrast, AWDDQN algorithm can overcome the shortcomings of DQN and DDQN algorithms. However, there are few applications of DDQN and AWDDQN algorithms in the field of voltage control in ADNs [24].

In this paper, an intelligent voltage control method based on AWDDQN algorithm is proposed for ADNs. The main contributions are as follows.

1) In the AWDDQN algorithm, dual weighted estimators are integrated into the DDQN algorithm to overcome the shortcomings of misestimation of action reward values by DQN and DDQN algorithms. The voltage control objective is incorporated into the designed action reward values to form the AWDDQN-based intelligent voltage controller for the ADN, where renewable power generations, loads, and adjustable resources of EVs are time-varying.

2) The capability of EVAs is quantified as adjustable resources for voltage control using the schedulable capacity approach of EVAs, ensuring that EV charging demand is satisfied after participating in dispatch.

3) Comprehensive comparisons of voltage control methods based on DQN, DDQN, AWDDQN, and traditional mixed-integer nonlinear program algorithms are presented, and the impact of inaccurate estimation results of action reward values on the performance of DQN and DDQN algo-

rithms is compared with the AWDDQN algorithm.

The remainder of this paper is organized as follows. Section II discusses the principle of the AWDDQN algorithm. The intelligent voltage control method based on AWDDQN algorithm is described in Section III. The effectiveness of the method is verified using digital simulation in Section IV, and the conclusions are drawn in Section V.

II. Principle of AWDDQN Algorithm

A. DQN Algorithm

An agent selects an action to act on the environment according to a certain policy in a state of the environment. Then, the environment gives a reward for the action and steps to the next state. This process is usually described in reinforcement learning as a Markov decision process (MDP) model, represented by a tuple $\langle S, A, P, r, \gamma \rangle$, where S is the state space, A is the action space, P is the state transition probability, r is the reward, and γ is the discount rate. One of the representative methods to solve the MDP model is the Q-learning algorithm, in which the expected value of function $Q(s,a)$ for action a taken for the state s in an episode needs to be evaluated. According to the Bellman optimality equation, this function is described as:

$$Q(s,a)=E\left(r+\gamma P(s,s')\max_{a'}Q(s',a')\right) \tag{1}$$

where $E(\cdot)$ is the expected value; $P(s,s')$ is the probability of reaching state s' from state s; and a' is the action in state s'.

The traditional Q-learning algorithm stores the action reward values of each state-action pair in a table and iteratively updates them continuously. However, storing and iterating the values of all state-action pairs are difficult when the state space is very large. To solve this problem, Q-learning algorithm is combined with deep learning to form the DQN algorithm [23]. Using neural networks, DQN algorithm learns to fit the true reward function $Q(s,a)$ with a parameterized approximate value function $Q(s,a|\theta)$, where θ is the parameter of $Q(s,a|\theta)$.

The structure of DQN is shown in Fig. 1, which is a forward neural network containing one input layer, one output layer, and multiple hidden layers.

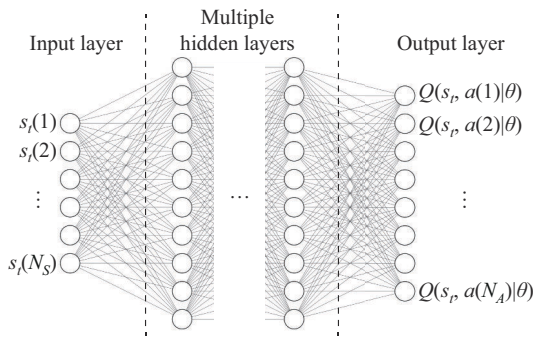

Fig. 1. Structure of DQN.

Its input is the current state s_t ($s_t \in S$) at current time t, and the number of inputs is the dimensions of state space elements N_S. Its output contains all the approximate action reward values of functions $Q(s_t,a|\theta)$ ($a \in A$) in that state s_t, and the dimension of outputs is that of action space N_A. The agent selects the action with the largest reward value in the output layer.

The loss function in DQN is expressed as:

$$L_{DQN}(\theta_i)=E(y_i^{DQN}-Q(s,a|\theta_i))^2 \tag{2}$$

where y_i^{DQN} and θ_i are the target function and the parameter of $Q(s,a|\theta)$ at iteration i, respectively. The target function is defined as:

$$y_i^{DQN}=r_i+\gamma Q(s',a^*_{DQN}|\theta_i^-) \tag{3}$$

where $Q(s',a^*_{DQN}|\theta_i^-)$ is the maximum value in the state s'; r_i is the reward at iteration i; θ_i^- is the parameter; and a^*_{DQN} is the action in the DQN algorithm with the maximum approximate value in the state s', as shown in (4).

$$a^*_{DQN}=\arg\max_{a'}Q(s',a'|\theta_i^-) \tag{4}$$

B. DDQN Algorithm

From (3) and (4), we can observe that the action a^*_{DQN} in DQN algorithm is selected and evaluated by the function $Q(s',a'|\theta_i^-)$. This makes it more likely to select overestimated values, resulting in overestimation of action reward values. To address this problem, the target function in DDQN algorithm y_i^{DDQN} is modified as:

$$y_i^{DDQN}=r_i+\gamma Q(s',a^*_{DDQN}|\theta_i^-) \tag{5}$$

where a^*_{DDQN} is the action with the maximum approximate value in the state s', as shown in (6).

$$a^*_{DDQN}=\arg\max_{a'}Q(s',a'|\theta_i) \tag{6}$$

From (5) and (6), we can observe that the action a^*_{DDQN} is selected by $Q(s',a'|\theta_i)$ and evaluated by $Q(s',a'|\theta_i^-)$ in DDQN algorithm. The probability of the action overestimated is greatly reduced by the design of those dual estimators.

However, the separation of selection and evaluation in DDQN algorithm sometimes creates underestimation problems, especially in environments with high stochasticity and uncertainty [25].

C. AWDDQN Algorithm

To overcome the above problems, the AWDDQN algorithm [26] is proposed in this paper for intelligent voltage control method in ADNs.

The structure of AWDDQN algorithm is similar to that of DQN algorithm, except the loss function and the target function. The loss function of the neural network in AWDDQN algorithm can be expressed as:

$$L(\theta_i)=E(y_i^{AWD}-Q(s,a|\theta_i))^2 \tag{7}$$

where y_i^{AWD} is the target function in AWDDQN algorithm. Dual weighted estimators are used in the target function, as shown in (8).

$$y_i^{AWD}=r_i+\gamma\left(\frac{\beta}{F}\sum_{f=i-F+1}^{i}Q(s',a^*|\theta_f)+\frac{1-\beta}{F}\sum_{f=i-F+1}^{i}Q(s',a^*|\theta_f^-)\right) \tag{8}$$

$$a^*=\arg\max_{a'}Q(s',a'|\theta_i) \tag{9}$$

where a^* is the action with the maximum approximate value

in the state s'; F is the memorized number of steps; θ_f and θ_f^- are the parameters of the approximate value functions of the online network and the target network at the f^{th} action in the past, respectively; and β is the weight, as shown in (10).

$$\beta = \frac{\left| Q(s', a^*|\theta_i^-) - Q(s', a_L|\theta_i^-) \right|}{c + \left| Q(s', a^*|\theta_i^-) - Q(s', a_L|\theta_i^-) \right|} \tag{10}$$

$$a_L = \arg\min_{a'} Q(s', a'|\theta_i) \tag{11}$$

where a_L is the action with the minimum approximate value in the state s'; and c is the hyperparameter for adjusting weight values.

After the loss function is established, the stochastic gradient descent algorithm is used to update the network parameter θ, as shown in (12).

$$\theta_{i+1} = \theta_i - \eta \nabla(L(\theta_i)) \tag{12}$$

where η is the gradient descent rate.

From (8), it can be observed that the action selection and evaluation are separated by the AWDDQN algorithm, thus avoiding the possible overestimation problem of action reward values of DQN algorithm. Therefore, it averages the F action estimation results learned in the past as the result of the target value, thus avoiding the possible misestimation of action reward values and the stochastic environment instability problem of DDQN algorithm. Using these dual weighted estimators and averaged weighted processing, better training stability and performance are obtained, thus achieving improved decision accuracy.

III. Intelligent Voltage Control Method Based on AWDDQN Algorithm

In the proposed method, the node voltages and the EVA and adjustable resource states constitute the state space, the active and reactive outputs of resources constitute the action space, and the weighted voltage fluctuations constitute the reward. This enables the agent to learn the most favorable output action for voltage control under different states.

A. Voltage Control Model for ADNs

The objective function for the voltage control can be expressed as:

$$\min \sum_{t=1}^{T} \left(\frac{1}{N} \sum_{n=1}^{N} t_s (U_{n,t} - U_{base})^2 \right) \tag{13}$$

where $U_{n,t}$ is the n^{th} node voltage at time t; N is the total number of nodes; T is the dispatch time horizon; t_s is the step interval; and U_{base} is the base voltage, usually set to be 1.0 p.u..

The constraints are as follows:

$$U_{\min} \le U_{n,t} \le U_{\max} \tag{14}$$

$$Q_{j,\min} \le Q_{j,t} \le Q_{j,\max} \tag{15}$$

$$P_{j,t,\min} \le P_{j,t} \le P_{j,t,\max} \tag{16}$$

where $U_{\max}$ and $U_{\min}$ are the allowable upper and lower voltage limits, respectively; $P_{j,t}$ and $Q_{j,t}$ are the active and reactive power of the j^{th} adjustable resource at time t, respectively; $Q_{j,\min}$ and $Q_{j,\max}$ are the minimum and maximum values of the adjustable reactive power of the j^{th} adjustable resource, respectively; and $P_{j,t,\min}$ and $P_{j,t,\max}$ are the minimum and maximum values of the adjustable active power of the j^{th} adjustable resource at time t, respectively.

In this paper, the adjustable reactive power resources considered are static reactive compensators with constant upper and lower limits. The adjustable active power resources considered are aggregated EVs, whose adjustable upper and lower limits are time-varying.

B. Schedulable Capacity of EVAs

The capability of EVAs as adjustable resources is measured by the schedulable capacity (EVSC), which is the bidirectional energy and power exchanged by an EVA with the grid at time t without affecting the future use of the EVs [27]. It includes the schedulable charging capacity (SCC), schedulable charging power (SCP), schedulable discharging capacity (SDC), and schedulable discharging power (SDP).

The calculations of SCC, SCP, SDC, and SDP are shown in (17)-(20).

$$SCC_{d,t} = C_d (SOC_{d,s} - SOC_{d,t}) + \eta_c (t + t_s - t_{d,s}) P_{d,c,\max} \tag{17}$$

$$SCP_{d,t} = \min(SCC_{d,t}/t_s \eta_c, P_{d,c,\max}) \tag{18}$$

$$SDC_{d,t} = C_d (SOC_{d,t} - SOC_d^{\min}) + (t_{d,e} - t - t_s) P_{d,c,\max}/\eta_d \tag{19}$$

$$SDP_{d,t} = \min(SDC_{d,t} \eta_d / t_s, P_{d,d,\max}) \tag{20}$$

where C_d is the battery capacity of the d^{th} EV; $SCC_{d,t}$, $SDC_{d,t}$, $SCP_{d,t}$, and $SDP_{d,t}$ are the SCC, SDC, SCP, and SDP of the d^{th} EV at time t, respectively; $P_{d,c,\max}$ and $P_{d,d,\max}$ are the maximum charging and discharging power of the d^{th} EV, respectively; η_c and η_d are the charging and discharging efficiencies, respectively; $t_{d,s}$ and $t_{d,e}$ are the excepted arriving and leaving time, respectively; t_s is the dispatch step; $SOC_{d,s}$ is the initial SOC of the d^{th} EV on arrival; and $SOC_{d,t}$ and $SOC_d^{\min}$ are the SOC of the d^{th} EV at time t and the minimum required SOC when leaving, respectively.

The SCC, SDC, SCP, or SDP of an EVA is the sum of the SCC, SDC, SCP, or SDP of EVs connected to this aggregator at time t, as in (21)-(24).

$$SCC_{l,t} = \sum_{d=1}^{N_{l,t}} SCC_{d,t} \tag{21}$$

$$SDC_{l,t} = \sum_{d=1}^{N_{l,t}} SDC_{d,t} \tag{22}$$

$$SCP_{l,t} = \sum_{d=1}^{N_{l,t}} SCP_{d,t} \tag{23}$$

$$SDP_{l,t} = \sum_{d=1}^{N_{l,t}} SDP_{d,t} \tag{24}$$

where $N_{l,t}$ is the total number of EVs connected to the l^{th} EVA at time t.

Equations (17) - (24) give a quantitative representation of the schedulable capacity and power range of EVs to participate in the voltage control in ADN without affecting the future use of EVs [28], [29]. The maximum adjustable active power of an EVA is the value of its SCP, and the minimum adjustable active power is the negative value of its SCP, i.e., $P_{j,t,\max} = SCP_{j,t}$ and $P_{j,t,\min} = -SDP_{j,t}$.

C. MDP Model

In reinforcement learning, the objective function and constraints of voltage control are normalized to form an MDP model.

1) State Space

In the proposed method, the state space S is the set of all states, and the state set s_t is the set of states of node voltages and adjustable resources of the ADN at time t, as shown in (25).

$$s_t=\{U_{i,t},\ldots,SCC_{l,t},\ldots,SDC_{l,t},\ldots,SCP_{l,t},\ldots,SDP_{l,t},\ldots,O_{j,t}\} \tag{25}$$

where $O_{j,t}$ is the output active or reactive power of the j^{th} ($j=1,2,\ldots,J$) adjustable resource at time t.

These data can be obtained by direct state measurement from the system devices in real systems. In this paper, the states are obtained by power flow calculations after adopting actions to simulate the operation of a real system. Specifically, the state space is divided into three parts, i.e., all node voltage values $U_{i,t}$, the EVSC of all dispatchable EVAs, and the output power of all adjustable resources $O_{j,t}$. If the number of adjustable EVAs is denoted by L, the dimension of state space N_S is equal to $N+4L+J$.

2) Action Space

The action space A includes all output cases of all adjustable resources. In this paper, the output cases are represented by the discretized outputs of the adjustable resources. Specifically, the action set a_t set for the state set s_t is a certain output action set of all adjustable resources, as shown in (26).

$$\begin{cases} a_t=\{a_{1,t}^{(k)},a_{2,t}^{(k)},\ldots,a_{J,t}^{(k)}\} \\ a_{j,t}^{(k)}=k \end{cases} \tag{26}$$

where $a_{j,t}^{(k)}$ is the k^{th} ($k=0,1,\ldots,K-1$) action of the j^{th} ($j=1,2,\ldots,J$) adjustable resource at time t, and K is the total number of actions per adjustable resource. Since the output cases of each adjustable resource are independent, the number of all possible cases is K^J, i.e., $N_A=K^J$.

The set of actions can be converted into the output power of all adjustable resources, thus allowing the action space and the state space to be linked, with the conversion formula, as shown in (27).

$$O_{j,t}=\begin{cases} Q_{j,\min}+\dfrac{Q_{j,\max}-Q_{j,\min}}{K-1}a_{j,t}^{(k)} & j \text{ is reactive unit} \\ P_{j,t,\min}+\dfrac{P_{j,t,\max}-P_{j,t,\min}}{K-1}a_{j,t}^{(k)} & j \text{ is active unit} \end{cases} \tag{27}$$

As can be observed from the design of the action space, the output ranges of the adjustable resources should not exceed their limits so that the constraints (15) and (16) are satisfied.

In addition, the action for EVs will be weighted according to the EVSC values of each EV, as shown in (28).

$$P_{d,t}=\begin{cases} O_{j,t}\dfrac{SCP_{d,t}}{P_{j,t,\max}} & O_{j,t}\geq 0 \\ O_{j,t}\dfrac{-SDP_{d,t}}{P_{j,t,\min}} & O_{j,t}<0 \end{cases} \tag{28}$$

where $P_{d,t}$ is the dispatched charging power of the d^{th} EV at time t.

3) Reward

The reward r_t obtained by the agent after selecting an action according to the node voltage states at time t is designed to have a negative value, thus, a higher reward means a smaller voltage volatility rate. The objective function can be obtained by accumulating the rewards, and this design makes the objective function (13) incorporated into the rewards, as shown in (29) and (30).

$$r_t=-\frac{1}{N}\sum_{i=1}^{N}\lambda_i t_s(U_{n,t+1}-U_{base})^2 \tag{29}$$

$$\lambda_i=\begin{cases} 1 & 0<\left|U_{n,t+1}-U_{base}\right|\leq 0.01 \\ 5 & 0.01<\left|U_{n,t+1}-U_{base}\right|\leq 0.03 \\ 10 & 0.03<\left|U_{n,t+1}-U_{base}\right|\leq 0.05 \\ 50 & \left|U_{n,t+1}-U_{base}\right|>0.05 \end{cases} \tag{30}$$

where λ_i is the penalty factor.

As can be observed from (29) and (30), by introducing the penalty factor λ_i, for constraint (14), a certain relaxation is achieved, and the penalty factor is higher for the nodes with larger voltage deviations. This allows the agent to prioritize the scheduling of nodes with larger voltage deviations in resource-limited scenarios to prevent the voltage from exceeding limits, and thus satisfy constraint (14).

D. Dynamic ε-greedy Strategy Design

A dynamic ε-greedy strategy is used during the agent training. At each iteration, the agent will select a random action with probability ε or select the action with the maximum reward with probability $1-\varepsilon$, while the value of ε changes. In this paper, a dynamic ε-greedy strategy combined with a simulated annealing method is used [30] to select actions in training.

$$\varepsilon_e=\exp\left(\frac{Q(s,a_r|\theta_i)-\max Q(s,a|\theta_i)}{\delta^e T_0}\right) \tag{31}$$

where ε_e is the ε value at the e^{th} episode; δ is the cooling down factor, which is a constant greater than 0 and lower than 1; T_0 is the initial temperature of simulated annealing; and a_r is a uniformly and randomly selected action in state s.

With the dynamic ε-greedy strategy combined with simulated annealing method, the goal of exploring first and then converging is achieved.

E. AWDDQN-based Intelligent Voltage Control Processes

The AWDDQN-based intelligent voltage control processes include the offline training and online voltage control processes, as shown in Algorithm 1. The purpose of offline training process is to teach the AWDDQN-based intelligent voltage control agent how to pick the optimal action for different states.

The AWDDQN-based intelligent voltage control agent is used as an online agent to control the real-time voltages of the ADN after the offline training is completed. The online voltage control process using the AWDDQN-based intelligent voltage control agent is shown in Fig. 2.

Algorithm 1

Input: the maximum iteration I, memory buffer size M, mini-batch size B, number of steps in an episode T, number of copy steps C, state s_0, c, γ, F, η, K, δ, and T_0
Initialize: neural network structure, memory replay buffer, parameters θ_0 and θ_0^-
For $i = 1$ to I, **do**
  Initialize state s_0
  For $t = 0$ to T, **do**
    Select action a_t based on the dynamic ε-greedy strategy with the output power of adjustable resources by (26)-(28)
    Change distribution network and EVs due to the power output of adjustable resources
    Obtain the new state s_{t+1} by (25)
    Calculate the reward r_t by (29) and (30)
    Store the tuple $\{s_t, a_t, s_{t+1}, r_t\}$ in the memory replay buffer
    if memory replay buffer is full
      Delete the earliest tuple
    end if
    Sample a mini-batch of B tuples from the memory, and replay buffer to the target and online network
    Obtain $Q(s, a|\theta_i)$ from the online network
    Calculate y_i^{AWD} of the target network by (8)
    Calculate $L(\theta_i)$ by (7)
    Update θ_{i+1} by (12)
    Input θ_{i+1} to the online network
    if $Mod(i,C) = 0$
      Copy θ_{i+1} to θ_i^-
    end if
  end for
end for
Output θ

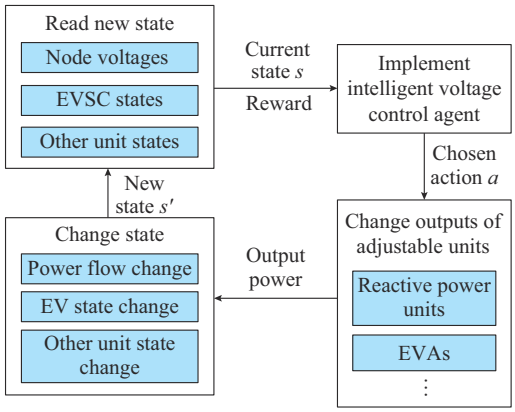

Fig. 2. Online voltage control process using AWDDQN-based intelligent voltage control agent.

The trained AWDDQN-based intelligent voltage control agent can decide the optimal output actions of the adjustable resources based on the system state information formed by node voltages, EVSC states, and other unit states.

In practice, the control procedure of EVs consists of two-level hierarchical control. In the upper-level control, the distribution system operator equipped with AWDDQN-based intelligent voltage control agent achieves voltage optimization by coordinating the schedulable capacity needed among the EVAs. In the lower-level control, each EVA coordinates the control of EV charging and discharging behaviors according to the negotiated schedulable capacity in the upper-level control. After receiving the action signal in the upper-level control, the EVAs coordinate the power of EVs by (28) in the lower-level control. Similarly, each reactive unit changes its output power so that the action is completed. In this way, the ADN voltages, the EVSC of EVAs, and the output power of other adjustable resources are updated to new states. Thus, the AWDDQN-based intelligent voltage control agent continuously selects the actions based on the new states. During the online control, the AWDDQN-based intelligent voltage control agent can handle the time-varying environment in real time through observing the states measured and the rewards received after the actual control actions.

IV. Simulation and Result Analysis

A. Test System and Parameter Settings

Modified versions of two typical distribution systems, i.e., the IEEE 33-bus and the IEEE 123-bus systems, are used as test cases. The settings of power resources and adjustable resources are shown in Table I.

TABLE I
Settings of Power Resources and Adjustable Resources

Testing system	Connected node	Resource type	Power setting
IEEE 33-bus	8, 25	PV generators	1.5 MW
	15	WT units	1.5 MW
	18, 23	EVAs	Shown in Fig. 3(c)
	30	Reactive power resources	[–0.5 Mvar, 0.5 Mvar]
IEEE 123-bus	20, 60, 86	PV generators	1.5 MW
	30, 68	WT units	1 MW
	29, 42, 52, 85	EVAs	Shown in Fig. 3(c)
	71, 109	Reactive power resources	[–0.5 Mvar, 0.5 Mvar]

The output profiles of WT units and PV generators are shown in Fig. 3(a), the load profiles are shown in Fig. 3(b), and the load base value is 3.5 MW in the modified IEEE 33-bus system and 5 MW in the modified IEEE 123-bus system, respectively. Besides, ±10% uniformly random fluctuations (represented by dashed line) are set in the RES output power and load to simulate realistic fluctuations. The EVSC profiles of EVAs are shown in Fig. 3(c). The base values of RESs, loads, and EVSC are also set with ±10% uniformly random fluctuations. Each EVA has 100 EVs with battery capacity of 40 kWh and charging and discharging power of 10 kW, and the charging and discharging efficiencies are 0.98. EVs are assumed to be charged at work place during the day, and some EVs with unmet charging demands during the day continue charging at home during the night.

From Fig. 3(c), it can be observed that EVs are connected to the grid after 07: 00, and their SCPs gradually increase.

With the increase of SOC, the SCPs of EVs decrease, and their SDPs increase. Most EVs are disconnected from the grid after 17:00, and both the schedulable charging power and discharging power of EVAs decrease. Some EVs are reconnected to the grid during the night, resulting in the increase and decrease of schedulable charging power and discharging power of EVAs again.

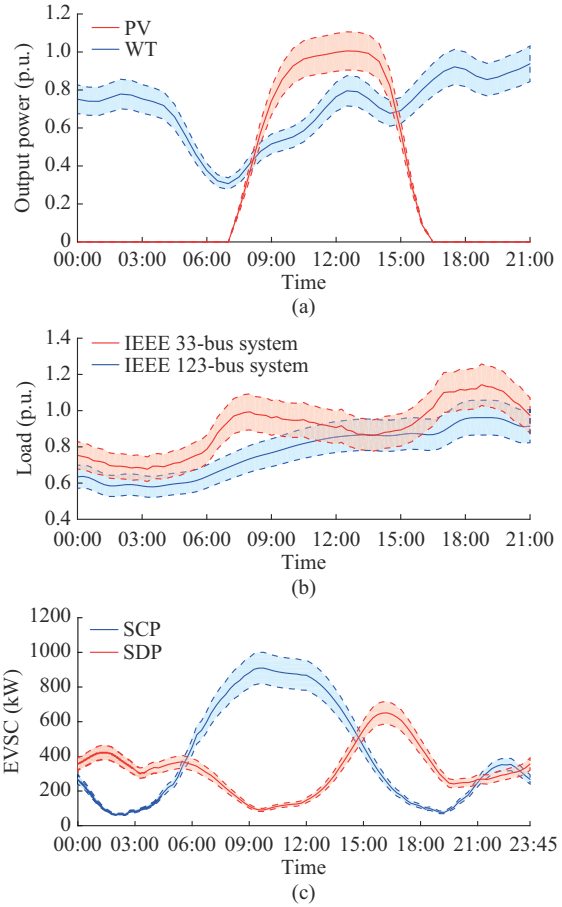

Fig. 3. System information of testing system. (a) Output power profiles of WT units and PV generators. (b) Load profiles. (c) EVSC profiles of EVAs.

In the IEEE 33-bus system, the number of nodes N is 33, the number of EVAs L is 2, and the number of adjustable resources J is 3, so the number of state space $N_S=N+4L+J=44$. The total number of adjustable actions per resource is $K=8$, so the number of action spaces is $N_A=K^J=512$. Therefore, the numbers of inputs and outputs to the neural network are 44 and 512, respectively. Similarly, in the IEEE 123-bus system, the numbers of inputs and outputs to the neural network are 140 and 32768, respectively.

B. Training Process Analysis

The 400-day data for output power of RESs, loads, and EVSCs are generated by using a Monte Carlo method which is set with ±10% uniformly random fluctuations from the base values. The step interval t_s is 15 min. The same method is used to generate the data in one day as the test set. The training parameters of the AWDDQN algorithm are given in Table II. And the hidden layer structures are {100, 100, 100} in IEEE 33-bus system, and {200, 200, 200} in IEEE 123-bus system. In DRL algorithm, the episode reward R (cumulative reward) obtained by an agent is used as evaluation criterion, which can be expressed as:

$$R=\sum_{t=0}^{T} r_t \tag{32}$$

In this paper, 24 hours are taken as one episode, and $T=96$. The calculation of r_t is shown in (29) and (30).

TABLE II
TRAINING PARAMETERS OF AWDDQN ALGORITHM

Parameter	IEEE 33-bus system	IEEE 123-bus system
Activation function	ReLU	ReLU
c	1	1
γ	0.99	0.99
η	0.001	0.001
F	5	4
M	10000	20000
B	200	200
C	200	200
I	288000	960000
δ	0.98	0.99
T_0	10^5	10^5

The simulation results of the voltage control methods with the DQN, DDQN, and AWDDQN algorithms are obtained and compared using a computer with 3.0 GHz Intel Core i7 CPU, 16 GB RAM, and GTX 2070 graphics card in the MATLAB environment. The training parameters of DQN and DDQN algorithms are the same as those of AWDDQN algorithm. The episode rewards of the three algorithms for the modified IEEE 33-bus system and IEEE 123-bus systems are shown in Fig. 4.

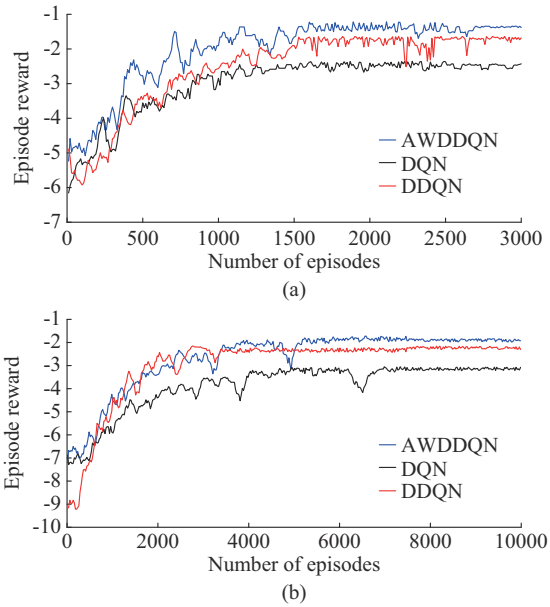

Fig. 4. Episode rewards of DQN, DDQN, and AWDDQN algorithms. (a) IEEE 33-bus system. (b) IEEE 123-bus system.

It can be observed from Fig. 4 that with the increase of training episode numbers, the three algorithms converge at similar speeds, all converging around 1500 episodes in the IEEE 33-bus system and 6000-7000 episodes in the IEEE 123-bus system. However, the final convergence trend shows that AWDDQN algorithm has the highest episode rewards among the three algorithms, which indicates that the AWDDQN algorithm can better evaluate the action reward values, as will be depicted in Section IV-C as well. In addition, it can be observed that the action reward values of all algorithms fluctuate highly at the early stage of training and gradually fluctuate less and converge. This indicates that the dynamic ε-greedy strategy has a better exploration at the early stage and a better stability at the later stage, which can achieve the goal of exploration first and convergence later.

C. Effect of K Value on Training

In the action space, K is the total number of actions per adjustable resource, i.e., the adjustable range of each adjustable resource is divided equally into K values, and the larger K is, the more precisely the resources can be adjusted, and vice versa. Besides, from (9), we can observe that the action a^* is obtained by traversing all actions a', thus the increase of dimensions of the action space delays the selection of the action. The number of dimensions in the action space is K^J, i. e., the number of outputs of AWDDQN algorithm is K^J, and its training speed is bound to become slower as K increases. Therefore, the investigation results of the effect of K in the IEEE 33-bus system are provided next.

Figure 5 shows the convergence process of AWDDQN algorithm, for K set to be 6, 8, and 10, corresponding to action space dimensions of 216, 512, and 1000 in the IEEE 33-bus system. It can be observed that as K increases, the convergence speed decreases, with $K=6$ converging around 1000 episodes, and with $K=10$ converging at 4000-5000 episodes. The results with the smallest convergence values at $K=6$ indicate that the action is not precise enough and the final results are poor, whereas the performances at $K=8$ and $K=10$ are similar. In addition, results with $K>10$ show that the improvement is not significant, while the time of offline training increases greatly. Therefore, the value of parameter K is finally selected as 8.

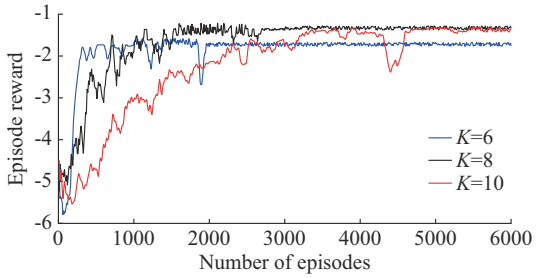

Fig. 5. Training process of AWDDQN algorithm with different K values in IEEE 33-bus system.

D. Simulation Results and Discussions

According to the above setting for the modified IEEE 33-bus and 123-bus systems, simulations are performed using the voltage control methods equipped with the DQN, DDQN, and AWDDQN algorithms, respectively. In addition, to demonstrate the advantages of the AWDDQN algorithm over traditional algorithms, the algorithm in [31] is also applied, in which the voltage optimization is formulated as a mixed-integer nonlinear program (MINLP) and can be solved by commercial solvers in MATLAB.

Figure 6 shows the node voltages without control throughout the day.

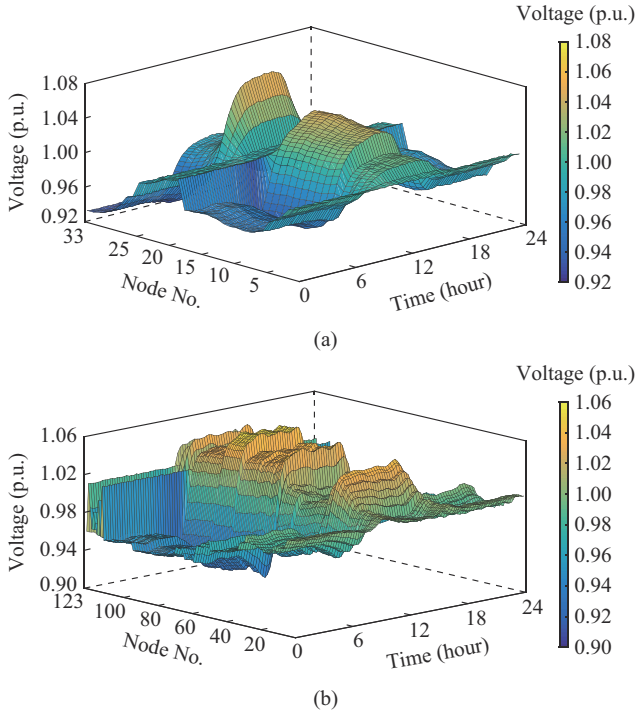

Fig. 6. Node voltages without control. (a) Voltages of IEEE 33-bus system. (b) Voltages of IEEE 123-bus system.

The reasonable range of voltage is [0.95, 1.05]p.u.. In the IEEE 33-bus system, most node voltages rise significantly at noon due to the increased output power of PVs, with a significant voltage over the upper limit near node 25. At night, due to the absence of PV power and increase of loads, there is a significant voltage drop at the end nodes such as node 33, and the voltage drop even below the lower limit. Similarly, there are node voltages outside limits in the IEEE 123-bus system, such as the voltage of node 86.

Figures 7 and 8 show the voltage control results obtained by the voltage control methods with the MINLP, DQN, DDQN, and AWDDQN algorithms for the IEEE 33-bus and 123-bus systems, respectively. In the IEEE 33-bus system, it can be observed that all methods can improve the voltage distribution, and the fluctuation ranges of node voltages become narrower. The voltage control methods with the AWDDQN, DDQN, and DQN algorithms have smoother voltage profiles in the control horizon. This indicates that DRL algorithms lead the voltage control agent to intelligently choose an effective output power action according to the current state. For example, using AWDDQN algorithm, when the PV power output reaches the maximum at noon, the agent effectively controls the EVs to absorb active power or controls

the reactive power resources to absorb reactive power to prevent overvoltage. At night, the agent controls the reactive power resources to achieve an optimal reactive power output or controls the discharge of EVs to improve end node voltages. Similar results are obtained in the IEEE 123-bus system.

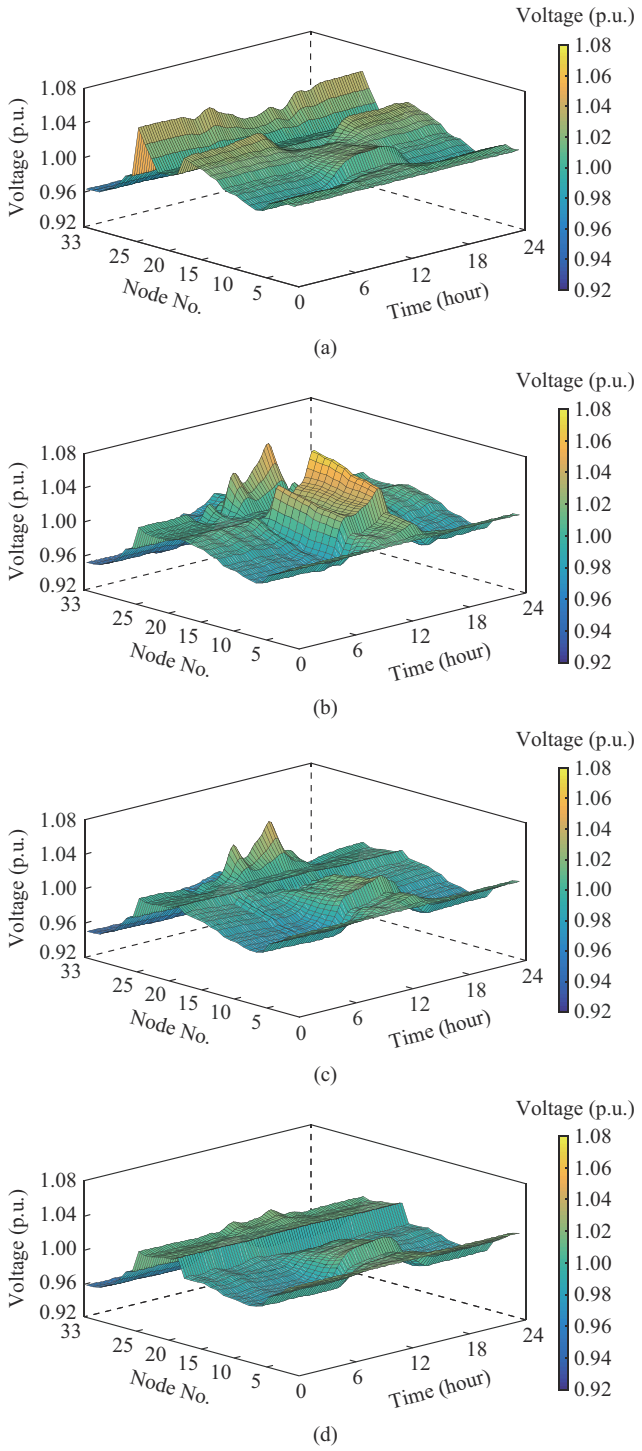

Fig. 7. Results of voltage control methods with different algorithms in IEEE 33-bus system. (a) MINLP algorithm. (b) DQN algorithm. (c) DDQN algorithm. (d) AWDDQN algorithm.

It can be observed from Figs. 7 and 8 that the overall control performance of the voltage control method with the AWDDQN algorithm is better than those of the methods with the DDQN and DQN algorithms. The AWDDQN algorithm mitigates the node voltage fluctuations more effectively compared with the DQN and DDQN algorithms. A detailed example of the optimal action results of the adjustable reactive power resources and the node voltage profiles using different DRL algorithms is shown in Fig. 9. These results indicate that the voltage control methods with the AWDDQN, DDQN, and DQN algorithms follow the same behaviors regardless of system scales.

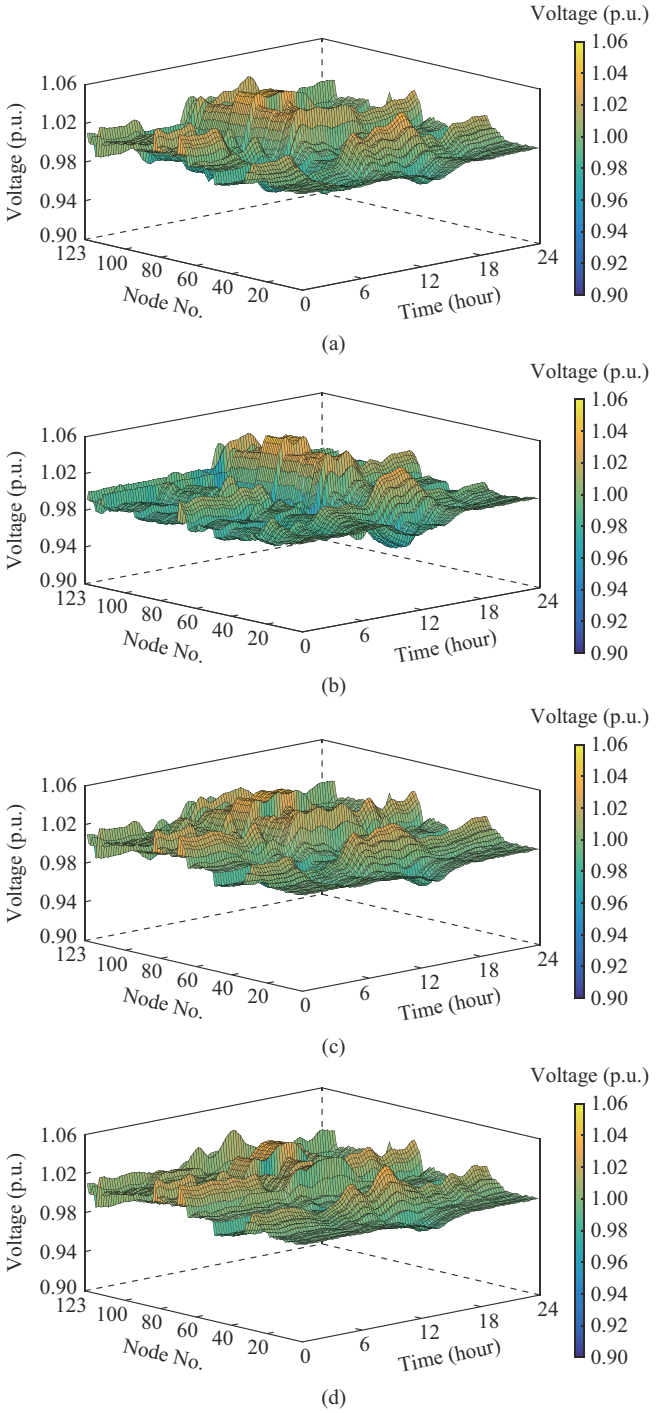

Fig. 8. Results of voltage control methods with different algorithms in IEEE 123-bus system. (a) MINLP algorithm. (b) DQN algorithm. (c) DDQN algorithm. (d) AWDDQN algorithm.

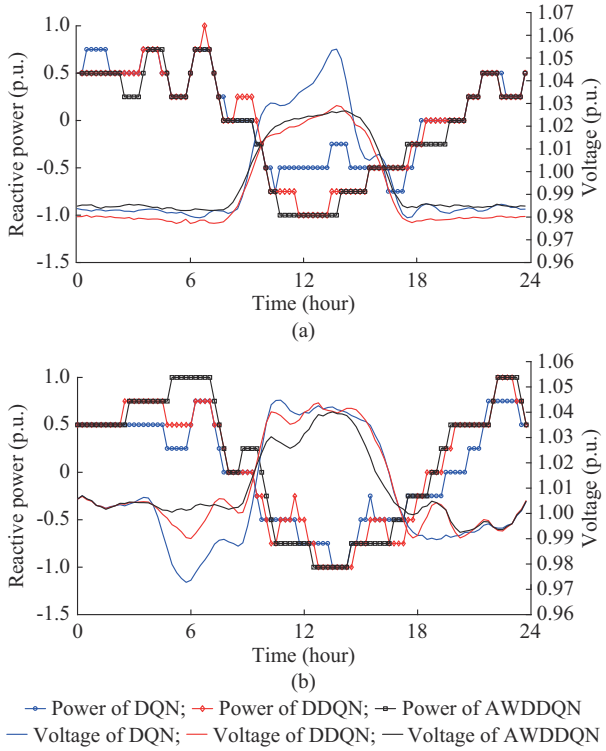

Fig. 9. Optimal action results of adjustable reactive power resources and node voltage profiles using voltage control methods using different DRL algorithms. (a) IEEE 33-bus system. (b) IEEE 123-bus system.

Figure 9(a) shows the action results of the reactive power resource 1 and voltages of node 8 using the voltage control methods with the DQN, DDQN, and AWDDQN algorithms in the IEEE 33-bus system. The reactive power absorption actions are taken during the daytime when the voltage is too high, while the reactive power output actions are taken during the night when the voltage is relatively low, thus supporting voltages. By comparing the results of the three voltage control methods, for example, during 10:00-15:00, it can be observed that the reactive power absorption actions of both DDQN and AWDDQN algorithms are higher than that of DQN algorithm, leading to better voltage control performance. This indicates that the DDQN and AWDDQN algorithms can alleviate misjudgments of the action reward value obtained by the DQN algorithm. Figure 9(b) shows the action results of the adjustable reactive power resource 1 and voltages of node 86 using the voltage control methods with the DQN, DDQN, and AWDDQN algorithms in the IEEE 123-bus system. By comparing the results of the voltage control methods based on the three RDL algorithms, for example, during 05:00-10:00, it can be observed that the reactive power output action of the AWDDQN algorithm reaches the maximum, whereas the DQN and DDQN algorithms request lower values, indicating that the AWDDQN algorithm can estimate the action reward value better than the DQN and DDQN algorithms.

The results of the two cases show that the control method with the AWDDQN algorithm can select the actions with higher rewards and thus obtain better control results compared with those with the DQN and DDQN algorithms. This proves that the control method with the AWDDQN algorithm is more accurate in evaluating the action reward values.

Table III shows the voltage control results, optimized objective function values, and calculation time with different algorithms.

TABLE III
VOLTAGE CONTROL RESULTS, OPTIMIZED OBJECTIVE FUNCTION VALUES, AND CALCULATION TIME WITH DIFFERENT ALGORITHMS

Testing system	Case	Objective function value	Voltage range	Calculation time (s)
IEEE 33-bus	Initial state	0.0248	[0.924, 1.072]	
	MINLP	0.0092	[0.962, 1.053]	213.45
	DQN	0.0103	[0.948, 1.060]	0.16
	DDQN	0.0080	[0.950, 1.051]	0.17
	AWDDQN	0.0072	[0.955, 1.020]	0.23
IEEE 123-bus	Initial state	0.0358	[0.919, 1.055]	
	MINLP	0.0046	[0.965, 1.045]	1026.45
	DQN	0.0057	[0.947, 1.048]	0.57
	DDQN	0.0030	[0.974, 1.044]	0.61
	AWDDQN	0.0025	[0.975, 1.040]	0.95

From Table III, it can be observed that in terms of voltage control results of the two systems, the voltage control method with the AWDDQN algorithm has the best performance among the four control methods, while the method with the DDQN algorithm has better performance than that with the DQN algorithm. In terms of the calculation time, voltage control methods with the AWDDQN, DDQN, and DQN algorithms require only 0.11%, 0.08%, and 0.08% of the time required by the method with MINLP algorithm for the IEEE 33-bus system, respectively, and 0.09%, 0.06%, and 0.06% of the time required by the method with the MINLP algorithm for the IEEE 123-bus system, respectively. Thus, voltage control methods with the DRL algorithms are superior to the method with the MINLP algorithm regarding execution speed. This is due to the large space of optimization variables and the large number of constraints. In both cases, the method with the AWDDQN algorithm requires slightly longer calculation time, while the control performance is improved compared with others. This is acceptable because the calculation time of the control methods with the DRL algorithms is very short in all cases.

Figure 10 shows the SOC curves of EVA 1 after scheduling by the voltage control method with the AWDDQN algorithm. In the IEEE 33-bus system, EVs arrive at the work place in the morning (06:00-10:00), and only a few EVs are charged during this period because there is no excess power supply in the morning and the system voltage is relatively low. Most of the EVs are charged at noon, since the grid voltage is high due to the high output of PVs. In the end, almost all EVs meet their charging demands, proving that the objective of distribution system operators and the EV individual demands can be satisfied. However, the charging demands of some EVs are not met, due to their short connection time. There is a similar trend in the SOC curves in the

IEEE 123-bus system.

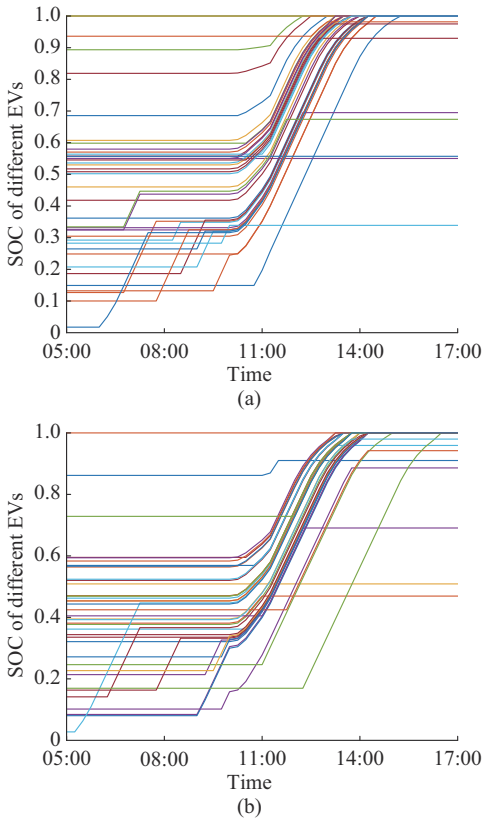

Fig. 10. SOC curves of EVA 1 after scheduling by voltage control method with AWDDQN algorithm. (a) IEEE 33-bus system. (b) IEEE 123-bus system.

V. Conclusion

In this paper, an intelligent voltage control method for ADNs based on AWDDQN algorithm is proposed. Using this method, the agent can intelligently control the adjustable active and reactive power resources according to the states of the ADN. The main conclusions are as follows.

1) The AWDDQN algorithm can intelligently coordinate and control the reactive power of resources and the active power of EVs for ADN voltage control, eliminating the overvoltage. The objective function is optimized from 0.0248 to 0.0072 in the IEEE 33-bus system and from 0.0358 to 0.0025 in the IEEE 123-bus system, respectively, without affecting the charging demands of EV users. These results indicate that the performance of the AWDDQN algorithm is unaffected by the scale of the systems.

2) The proposed method has a much faster speed than that with the traditional MINLP algorithms.

3) The simulation results for the IEEE 33-bus and IEEE 123-bus systems indicate the problems of overestimation in the voltage control method with the DQN algorithm and the underestimation in that with the DDQN algorithm.

4) The proposed method with the AWDDQN algorithm can overcome these shortcomings by introducing the average weighted estimators, resulting in better evaluation of the action reward values and better reward convergence values.

5) The complexity of the design of AWDDQN target increases the calculation time at acceptable levels.

References

[1] H. Zhou, S. Chen, J. Lai *et al.*, "Modeling and synchronization stability of low-voltage active distribution networks with large-scale distributed generations," *IEEE Access*, vol. 6, pp. 70989-71002, Nov. 2018.

[2] S. Xia, S. Bu, C. Wan *et al.*, "A fully distributed hierarchical control framework for coordinated operation of DERs in active distribution power networks," *IEEE Transactions on Power Systems*, vol. 34, no. 6, pp. 5184-5197, Nov. 2019.

[3] C. Sarimuthu, V. Ramachandaramurthy, K. Agileswari *et al.*, "A review on voltage control methods using on-load tap changer transformers for networks with renewable energy sources," *Renewable and Sustainable Energy Reviews*, vol. 62, no. 1, pp. 1154-1161, Sept. 2016.

[4] Z. Liu and X. Guo, "Control strategy optimization of voltage source converter connected to various types of AC systems," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 1, pp. 77-84, Jan. 2021.

[5] R. A. Jabr, "Power flow based volt/var optimization under uncertainty," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 5, pp. 1000-1006, Sept. 2021.

[6] H. Li, M. A. Azzouz, and A. A. Hamad, "Cooperative voltage control in MV distribution networks with electric vehicle charging stations and photovoltaic DGs," *IEEE Systems Journal*, vol. 15, no. 2, pp. 2989-3000, Jun. 2020.

[7] Y. Zheng, Y. Song, D. J. Hill *et al.*, "Online distributed MPC-based optimal scheduling for EV charging stations in distribution systems," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 2, pp. 638-649, Feb. 2019.

[8] M. Mazumder and S. Debbarma, "EV charging stations with a provision of V2G and voltage support in a distribution network," *IEEE Systems Journal*, vol. 15, no. 1, pp. 662-671, Mar. 2021.

[9] A. Ahmadian, B. Mohammadi-Ivatloo, and A. Elkamel, "A review on plug-in electric vehicles: introduction, current status, and load modeling techniques," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 3, pp. 412-425, May 2020.

[10] H. Patil and V. N. Kalkhambkar, "Grid integration of electric vehicles for economic benefits: a review," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 1, pp. 13-26, Jan. 2021.

[11] X. Sun and J. Qiu, "Hierarchical voltage control strategy in distribution networks considering customized charging navigation of electric vehicles," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 4752-4764, Nov. 2021.

[12] Y. Wang, T. John, and B. Xiong, "A two-level coordinated voltage control scheme of electric vehicle chargers in low-voltage distribution networks," *Electric Power Systems Research*, vol. 168, no. 1, pp. 218-227, Mar. 2018.

[13] Y. Liu and H. Liang, "A discounted stochastic multiplayer game approach for vehicle-to-grid voltage regulation," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 10, pp. 9647-9659, Oct. 2019.

[14] J. Hu, C. Ye, Y. Ding *et al.*, "A distributed MPC to exploit reactive power V2G for real-time voltage regulation in distribution networks," *IEEE Transactions on Smart Grid*, vol. 13, no. 1, pp. 576-588, Jan. 2022.

[15] Y. Zhang, X. Wang, J. Wang *et al.*, "Deep reinforcement learning based volt-var optimization in smart distribution systems," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 361-371, Jan. 2021.

[16] H. Diao, M. Yang, F. Chen *et al.*, "Reactive power and voltage optimization control approach of the regional power grid based on reinforcement learning theory," *Transactions of China Electrotechnical Society*, vol. 30, no. 12, pp. 408-414, Jun. 2015.

[17] D. Cao, W. Hu, J. Zhao *et al.*, "Reinforcement learning and its applications in modern power and energy systems: a review," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1029-1042, Dec. 2020.

[18] V. Mnih, K. Kavukcuoglu, D. Silver *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529-533, Feb. 2015.

[19] J. Shi, W. Zhou, N. Zhang *et al.*, "Deep reinforcement learning algorithm of voltage regulation in distribution network with energy storage system," *Electric Power Construction*, vol. 3, pp. 1-8, Mar. 2020.

[20] R. Diao, Z. Wang, D. Shi *et al.*, "Autonomous voltage control for grid operation using deep reinforcement learning," in *Proceeding of 2019 IEEE PES General Meeting (PESGM)*, Atlanta, USA, Aug. 2019, pp. 1-5.

[21] Q. Yang, G. Wang, A. Sadeghi *et al*., "Two-timescale voltage control in distribution grids using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2313-2323, May 2020.
[22] X. Sun and J. Qiu, "A customized voltage control strategy for electric vehicles in distribution networks with reinforcement learning method," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 10, pp. 6852-6863, Oct. 2021.
[23] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, pp. 1-13, Mar. 2016.
[24] O. Lukianykhin and T. Bogodorova, "Voltage control-based ancillary service using deep reinforcement learning" *Energies*, vol. 14, no. 8, pp. 1-22. Apr. 2021,
[25] Z. Zhang, Z. Pan, and M. J. Kochenderfer, "Weighted double Q-learning," in *Proceeding of International Joint Conference on Artificial Intelligence*, Melbourne, Australia, Aug. 2018, pp. 3455-3461.
[26] J. Wu, Q. Liu, S. Chen *et al*., "Averaged weighted double deep Q-network," *Journal of Computer Research and Development*, vol. 57, no. 3, pp. 576-589, Jun. 2020.
[27] H. Zhang, Z. Hu, Z. Xu *et al*., "Evaluation of achievable vehicle-to-grid capacity using aggregate PEV model," *IEEE Transactions on Power Systems*, vol. 32, no. 1, pp. 784-794, Jan. 2017.
[28] H. Liang, Z. Lee, and G. Li, "A calculation model of charge and discharge capacity of electric vehicle cluster based on trip chain," *IEEE Access*, vol. 8, pp. 142026-142042, Aug. 2020.
[29] F. D. Kanellos, "Optimal scheduling and real-time operation of distribution networks with high penetration of plug-in electric vehicles," *IEEE Systems Journal*, vol. 15, no. 3, pp. 3938-3947, Sept. 2021.
[30] R. Su, F. Wu, and J. Zhao, "Deep reinforcement learning method based on DDPG with simulated annealing for satellite attitude control system," in *Proceedings of* 2019 *Chinese Automation Congress (CAC)*, Hangzhou, China, Nov. 2019, pp. 390-395.
[31] Z. Wang, J. Wang, B. Chen *et al*., "MPC-based voltage/var optimization for distribution circuits with distributed generators and exponential load models," *IEEE Transactions on Smart Grid*, vol. 5, no. 5, pp. 2412-2420, Sept. 2014.

Yangyang Wang received the B.Sc. degree in electrical engineering and automation from Hefei University of Technology, Hefei, China, in 2014. He is currently pursuing a Ph. D. degree in electrical power system and automation at Hefei University of Technology. His research interests include EV integration into smart grid, machine learning in power system, and energy management system in microgrids.

Meiqin Mao received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Hefei University of Technology, Hefei, China, in 1983, 1988, and 2004, respectively. She is now a Professor with School of Electrical and Automation Engineering, Hefei University of Technology. She has published more than 200 journal/conference technical papers. She now serves as an Associate Editor for IEEE Journal of Emerging and Selected Topics in Power Electronics. Her research interests include renewable energy generation technology, distributed power generation and microgrids, and power electronics applied in power system.

Liuchen Chang received the B.Sc. degree from Northern Jiaotong University, Beijing, China, in 1982, the M. Sc. degree from the China Academy of Railway Sciences, Beijing, China, in 1984, and the Ph. D. degree from Queen's University, Kingston, Canada, in 1991. He is a Professor of electrical and computer engineering and NSERC Chair at the University of New Brunswick, Fredericton, Canada. He is a Fellow of Canadian Academy of Engineering. His research interests include distributed power generation, renewable energy, analysis and design of electrical machines, variable speed drives, power electronics, and EV traction system.

Nikos D. Hatziargyriou received the B.Sc. degree from National Technical University of Athens, Athens, Greece, in 1976, the M.Sc. and Ph.D. degrees from The University of Manchester, Manchester, UK, in 1979 and 1982, respectively. He is a Professor Emeritus in power systems at the National Technical University of Athens. He is Life Fellow Member of IEEE, past Chair of the Power System Dynamic Performance Committee (PSDPC) and past Editor in Chief (EiC) of the IEEE Transactions on Power Systems and currently EiC at Large for PES Transactions. His research interests include smart grids, microgrids, distributed and renewable energy sources, and power system security.