Reactive Power-Voltage Coordinated Control of Offshore Wind Farm Based on Multi-Agent Reinforcement Learning

Abstract: This paper proposes a distributed reactive power-voltage (Q-V) coordinated control approach based on multi-agent deep reinforcement learning algorithm. Firstly, the Q-V control problem is formulated as a Markov game where all wind turbines (WTs) on each feeder is modeled as an adaptive agent. Secondly, the policy networks of each agent are trained by the advanced multi-agent deep deterministic policy gradient method. Then, the trained policy networks are executed in a distributed manner to control the voltage. The proposed method can significantly reduce the requirements of communications and knowledge of system parameters. It also effectively deals with uncertainties and can provide online coordinated control only based on local information. The simulation results of connecting the wind farm with the IEEE 14-bus system demonstrate the effectiveness and benefits of the proposed approach.

Key words: Offshore Wind Farm Control, Reactive Power-Voltage Regulation, Multi-Agent Deep Reinforcement Learning, Distributed Network

I. Introduction

Offshore wind energy is one of the impactful ways to alleviate the current energy and environmental concerns [1.1-1.2]. However, due to offshore wind's uncertainties and volatility characteristics, its higher integration brings numerous technical challenges to the voltage control of offshore wind farms. Meanwhile, offshore wind farms are mainly connected to the grid by AC transmission. Due to the capacitive effect of the AC submarine cable, a large amount of charging power increases the voltage at the end of the cable [2.1, 2.2], which results in a lower voltage stability margin of the WT. In addition, it is difficult and costly to install reactive power compensation equipment in an offshore wind farm. Therefore, the Q-V coordinated control of WTs is essential and is more economical in an offshore wind farm [2.3].

The Q-V coordinated control method, based on OPF theory, determines the reactive power output of each WT according to the status of WTs in a wind farm. In [2.4, 2.5], an OPF model based on reactive power dispatch method is established, aiming to reduce systems power losses, including WTs and collector cable equipment. In [2.6], an OPF method is used to calculate the voltage reference of the pilot bus, then the total reactive power demand is determined through a PI control and dispatched proportionally to each WT. In [2.7], based on model predictive control (MPC), a coordinated optimal control model with different time scales for reactive power compensation equipment in a wind farm is established, aiming to coordinate different reactive power compensation equipment, reduce voltage deviation, and improve system reactive power margin.

In solving the OPF model, the particle swarm optimization algorithm is used to solve the optimal voltage regulation model in [2.10]. In [2.11], a sensitivity analysis combined with the MPC method is proposed to solve the wind farm OPF model. [2.12, 2.13] divided wind farms into clusters, and a distributed algorithm is used to solve to improve the speed. In [2.14], a deep learning intelligent voltage regulation framework was established, which used historical data to train the model and realized online response. Among the above methods, nonlinear solution methods have high solution accuracy, but it is not easy to guarantee real-time performance. The sensitivity-based linearization method improves the solution speed but cannot take the accuracy of the solution into account. The data-driven method relies on a large amount of historical data.

To sum up, though there have been researches on the model and solution of Q-V coordinated control for wind farms, two major problems remain: the solution time and accuracy are difficult to guarantee due to the nonlinear characteristics of the OPF model; the traditional machine learning method highly relies on real-world historical data.

Given this, this paper proposes a distributed Q-V control method for the offshore wind farm based on multi-agent deep reinforcement learning (MADRL). Firstly, a reactive optimal power flow (Q-OPF) model of the wind farm considering node voltage deviation is established. Secondly, based on the Q-OPF model, the Q-V control problem is formulated as a Markov game where the WTs on a feeder are modeled as an adaptive agent. Thirdly, the policy networks of each agent are trained by the advanced multi-agent deep deterministic policy gradient (MADDPG) method. Then, the trained policy networks are executed in a distributed manner to control the voltage. The simulation results using random output data of WTs show that the proposed method can effectively improve the voltage stability of wind farms without relying on historical data and has better model solution accuracy and speed performance than traditional methods.

The rest of the paper is organized as follows. In section II, the Q-V coordinated control model of the offshore wind farm is presented. Section III describes the proposed method. In section IV, the simulation results are illustrated in detail. Finally, Section V concludes this paper.

II. Problem Formulation

The structure of the wind farm is shown in Fig.1. The 5MW high-speed permanent magnet wind turbine is adopted. The output voltage of the WT is 0.69kV, which is boosted to 35kV by its transformer. All WTs are connected to a three-phase double winding transformer through a submarine cable and then connected to the onshore power grid through a submarine AC transmission cable. The lower part of Fig.1 shows the structure of the WT. The parallel full power converter is used, and the average power distribution control is adopted. The generator side converter realizes the active and reactive power control of the generator, and the grid side converter realizes the DC bus voltage control and reactive power (injected into the grid) control.

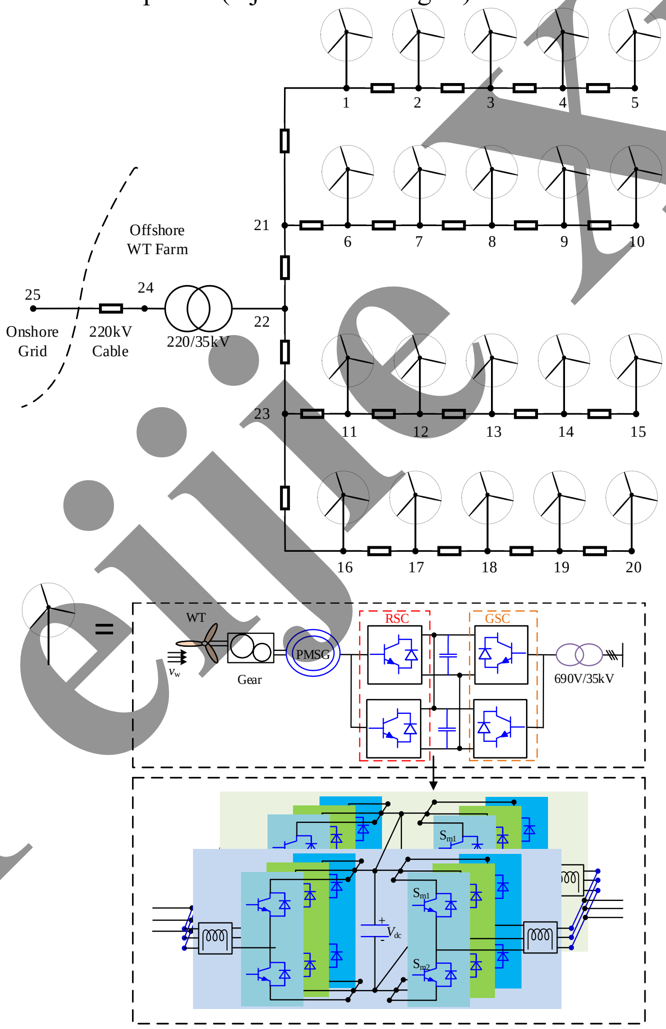

Fig. 1. The architecture of the target offshore wind farm

The Q-V coordinated control problem is formulated as follows:

1) Object

$$\min \sum (U_i - U_{i-ref})^2 \qquad i \in [1, N_s]$$

2) Equality constraint

$$\begin{cases} P_i = U_i \sum_{j=1}^{Ns} U_j (G_{ij} \cos\theta_{ij} + B_{ij} \sin\theta_{ij}) \\ Q_i = U_i \sum_{j=1}^{Ns} U_j (G_{ij} \sin\theta_{ij} - B_{ij} \cos\theta_{ij}) \end{cases} \qquad i, j \in [1, N_s], i \neq j$$

3) Inequality constraints

$$\begin{cases} Q_{i\min} \le Q_i \le Q_{i\max} \\ U_{i\min} \le U_i \le U_{i\max} \\ S_{i\min} \le S_i \le S_{i\max} \end{cases} \qquad i \in [1, N_s]$$

where (1) is the objective function to minimize the sum of the voltage deviation of each node; U_i and $U_{i\text{-}ref}$ donate the voltage and the rated voltage of the WT of node *i* in wind farm; (2) is the equality power flow constraints, where P_i and Q_i are active and reactive power that generated by the WT of node *i*; G_{ij} and B_{ij} are the real and imaginary part of admittance element between nodes *i* and *j*; θ_{ij} is the voltage phase difference between nodes *i* and *j*; (3) is inequality constraints, where S_i is apparent power that generated by the WT of node *i* and its lower and upper limits are donated by $S_{i\min}$ and $S_{i\max}$; similar to that, $Q_{i\min}$ and $Q_{i\max}$ are the lower and upper limits of the reactive power of the WT, $U_{i\min}$ and $U_{i\max}$ are the lower and upper limits of the voltage of the WT.

III. Proposed Multi-Agent Control for Wind Farm

The proposed distributed Q-V coordinated control is shown in Fig. 2, which is based on the multi-agent deep deterministic policy gradient (MADDPG) method. It contains three main steps, namely: 1) grouping the WTs on the same feeder; 2) formulating the decomposed sub-networks as agents in a Markov game; and 3) training the deep neural network (DNN) of the agents for coordinated voltage control via adapted MADDPG method.

A. Sub-network Grouping

In this work, we group the WTs on the same feeder as one agent, as they have a strong electrical connection on their active power, reactive power, and voltage. Meanwhile, they are close to each other from a perspective of space, and therefore, communication between them will also be more manageable. As shown in Fig. 2, the target wind farm was divided into four agents when executing Q-V control.

B. Markov Game in the Q-V Coordinated Control

In this section, we formulate the distributed Q-V coordinated control as a Markov Game (MG). In the MG, each sub-network is modeled as an adaptive agent, which makes control decisions based only on local information of the corresponding sub-network at each time step. The key components for an MG include state set S, action set A, and reward function R.

S: the state set S_t contains the states for all agents. For agent j, s_t^j denotes for its state at time step t, which also means the local observation of sub-network j. s_t^j includes $\{(U_t^i, P_t^i, \Delta P_t^i)\}$, where ΔP is the forward difference of active power, i is the index of the node that is located in sub-network j.

A: the action set A_t contains the actions for all agents. For agent j, the action at time step t, a_t^j is $\{Q_t^i\}$, where i is the index of the node that is located in sub-network j.

R: $r_t^j \in R_t$ is the immediate reward the agent j obtains after the action a_t^j is executed. In this context, all the agents share the same reward: $r_t = -\sum_{i=1}^{N_s} (U_i - U_{i-ref})^2$ which represents the total voltage deviation of all WTs at time step t; to lead the agents make a decision that meets the power limits, the overflow of the apparent power of each WT is added to the reward; therefore, the final reward function is $r_t = -\sum_{i=1}^{N_s}[w_i(U_i - U_{i-ref})^2 + k_i(S_i - S_{i\max})]$, where w_i and k_i are the weights to balance the importance of voltage deviation as power limits.

At time step t, agent j makes its decision a_t^j based on the local observation s_t^j of sub-network j. When all agents complete their actions, they obtain a shared reward r_t, and then the system transfers to the next state. This is an MG and the goal of each agent is to learn a policy, which maps its local observation s_t^j to action a_t^j in order to maximize the discounted cumulative reward from the current time-step onward, $\sum_{k=t}^{T}\gamma^{k-t}r_k$, where $\gamma \in [0,1)$ is the discount factor that balances the importance between the future and immediate reward.

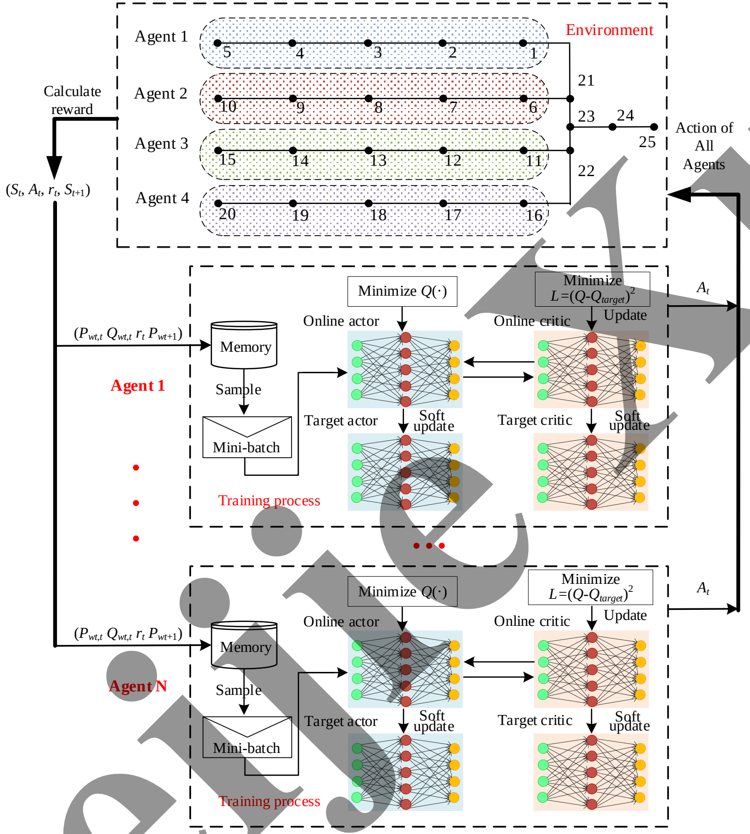

Fig. 2. The frame of the proposed MADDPG-based Q-V coordinated control

C. Adapted MADDPG Algorithm for Q-V Coordinated Control

The adapted MADDPG algorithm is developed to solve the MG in the Q-V coordinated control. Each sub-network is modeled as an agent, which is composed of the actor and critic DNNs. The actor, which is the policy network, maps the local observation s_t^j to action a_t^j. The critic maps global information (S_t, A_t) from all agents to a scalar, which is a judgment of action a_t^j considering the impact on other agents. The coordinated control strategy is achieved by adopting a centralized training framework, among which the actor and critic networks of each agent are trained against each other iteratively till the critic provides a suitable judgment and the actor can make decisions with reduced voltage deviation.

For the agent j, let the actor network be parameterized by μ_θ^j and the critic network be parameterized by Q_θ^j; therefore, we have $a_t^j = \mu_\theta^j(s_t^j)$ to be the decision made by the actor and $Q_\theta^j(s_t^j, a_t^1, ..., a_t^j, ..., a_t^N)$ to be the output of the critic network, where N is the number of agents. Meanwhile, the algorithm introduces a target actor-critic network μ'^j_θ and Q'^j_θ to the agent in order to prevent the unstableness of the training process. Also, each agent has a replay buffer, which is in charge of storing the transitions $(s_t^j, a_t^j, r_t^j, s_{t+1}^j)$. The mini-batch experiences are sampled at each time step to calculate the gradient and optimize the parameters of networks. This mechanism helps break the correlation between data and improves the stability of the training process.

Table I.

```
Algorithm: MADDPG for N agents in a wind farm
For each agent j, randomly initialize parameters of actor network $\mu_\theta^j$ and critic network $Q_\theta^j$
For each agent j, initialize parameters of the target network, $\mu'^j_\theta \leftarrow \mu_\theta^j, Q'^j_\theta \leftarrow Q_\theta^j$
for episode e = 1, 2, …, H do
    Initialize a random process $\varepsilon$ for action exploration
    For each agent j, receive initial state $s_0^j$
    for time step t = 1,2, …, T do
        For each agent j, select action $a_t^j = \mu_\theta^j(s_t^j) + \varepsilon$
        Execute actions $A_t = (a_t^1, ..., a_t^j, ..., a_t^N)$ and get observation $S_t$, reward $R_t$, and new state $S_{t+1}$
        Store the experience $(s_t^j, a_t^j, r_t^j, s_{t+1}^j)$ in the replay buffer D
        if t % learning_period == 0 do
            for agent j = 1,2, …, N do
                Randomly sample a mini-batch experience B from D
                Set $y_t^j = r_t^j + \gamma Q'^j_\theta(s_{t+1}^j, a'^1_{t+1}, ..., a'^j_{t+1}, ..., a'^N_{t+1})$, where $a'^j_{t+1} = \mu'^j_\theta(s_{t+1}^j)$
                Update critic network by minimizing the loss:
                $L(\theta_j) = \frac{1}{B}\sum(y_t^j - Q_\theta^j(s_t^j, a_t^1, ..., a_t^j, ..., a_t^N))^2$
                Update actor network using the mini-batch policy gradient:
                $\nabla_\theta J = \frac{1}{B}\sum \nabla_\theta \mu_\theta^j(s_t^j) \nabla_{a_j} Q_\theta^j(s_t^j, a_t^1, ..., a_t^j, ..., a_t^N)|_{a_j = \mu_\theta^j(s_t^j)}$
            end for
            Update target network: $\mu'^j_\theta \leftarrow \tau\mu_\theta^j + (1-\tau)\mu'^j_\theta$, $Q'^j_\theta \leftarrow \tau Q_\theta^j + (1-\tau)Q'^j_\theta$
        end if
    end for
end for
```

D. Real-time Voltage Control of the Proposed Approach

When the training process is completed, the parameters of DNN are fixed, and only the actor network of each agent is kept for real-time voltage regulation. Each agent is in charge of a sub-network. The real-time reactive power control scheme of the proposed approach is shown in Table II. The centralized critics augmented with information of other agents' policies during the training process help formulate coordinated strategies. The explicitly modeling of other agents' decision-making process allows each agent to provide decisions with better robustness to system dynamics based on local information only. This differentiates the existing DDPG-based works and allows us to deal with scalability issues in the presence of large-scale systems.

Table II

```
Algorithm: Real-time distributed Q-V control
For each agent j, load the parameters of actor network $\mu_\theta^j$
for time step t = 1, 2, …, T do
    for agent j = 1,2, …, N do
        Obtain the local observation $s_t^j$
        Calculate action: $a_t^j = \mu_\theta^j(s_t^j)$
    end for
    Concatenate the actions $A_t = (a_t^1, ..., a_t^j, ..., a_t^N)$
    executed action $A_t$
end for
```

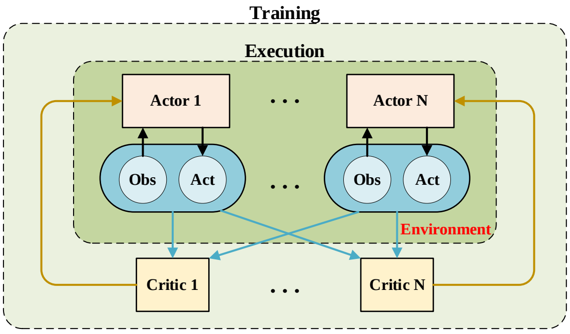

Fig. 3. Centralized training and decentralized execution of the proposed MADDPG-based Q-V coordinated control

IV Numerical Results

In this section, simulation results are provided to evaluate the performance of the proposed approach on the IEEE 14-bus system. The target voltage of WTs is set at 1.03 p.u. and then at 1.00 p.u. to illustrate the generality of the proposed MADDPG-based method.

A. Simulation Setup

Firstly, we define the average voltage deviation (AVD) $AVD = \frac{1}{N_s}\sum_{i=1}^{N_s}|U_i - U_{ref}|$ to be the evaluation index of the policy network, where U_{ref} denotes the target voltage.

To simulate more realistic scenarios, as shown in Fig. 4, we connected the wind farm to node 4 in the IEEE 14-bus system. Moreover, real-world active power data of the target wind farm are used. These active power data have 440 steps, including a power jump at around the 140th step. The proposed approach is implemented in Python with PaddlePaddle. The power flow is calculated by Pypower. Baidu AI Studio provides the programming platform and deep learning frame.

Table III. Parameters of the Target Offshore Wind Farm

Parameters	Value
35kVCable	R=0.0754 Ω/km; L=0.3365 mH/km; C=0.1805 mF/km
35kV/220kV Transformer	R=0.005 p.u.; X=0.12 p.u.
220kVCable	R=0.0221 Ω/km; L=0.446 mH/km; C=0.155 mF/km

Table IV. Parameters Settings of the Proposed Method

Parameters	Value
Batch size	32
Replay buffer size	100000
Discount factor	0.75
Soft update coefficient	0.001
Policy update period	3
Weights of the voltage on a feeder	1:0.2:1.8
Learning rate of actor network	0.001
Learning rate of critic network	0.001

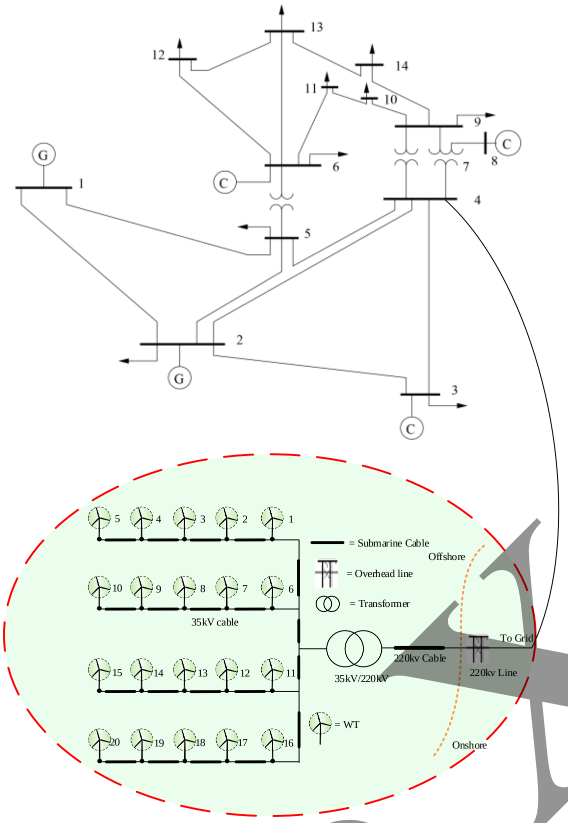

Fig. 4. Modified IEEE 14-bus system with the target wind farm.

B. Performance Evaluation

The proposed approach is trained for 50000 epochs on the training data to learn the coordinated control strategy for voltage regulation. The convergence curve of the cumulative reward is plotted in Fig. 5. It can be observed that the proposed approach could not make balanced decisions at the beginning of the training procedure and therefore achieves low reward. With the training process going on, the reward increases significantly and finally converges around -0.5 with minor fluctuations, illustrating that the proposed method can learn the coordinated control strategy.

Fig. 6. shows the distribution of the WTs' voltage with the target at 1.03 p.u., all of them are in a range from 1.029 p.u. to 1.031 p.u., with the AVD at 3.6×10^{-4} p.u..

Fig. 7 shows the voltage of five WTs on the first feeder, whose policy network is trained by DDPG and proposed MADDPG-based method respectively. In the upper part of Fig. 7, the voltage of five WTs experience a jump due to the jump of their active power; apparently, the policy network trained by DDPG cannot handle such sudden change at a short time. In the lower part of Fig. 7, in the contrast, the policy network trained by proposed MADDPG-based method can effectively reduce the potential voltage jump.

Fig.8 shows the reactive power of each WT on the first feeder, which is also the action of the first agent in the MG; it shows that every dimensions of the action are changing actively to minimize voltage deviation.

To validate the generality of the proposed MADDPG-based method, we set the reference voltage at 1.00 p.u. and train the agents using the proposed MADDPG-based method. As it shown in Fig. 9 all of them are in a range from 0.999 p.u. to 1.001 p.u., with the AVD at 3.3×10^{-4} p.u..

D. Performance Improvement of Proposed Method

As shown in Table V and Fig. 10, when the target voltage was set at 1.03 p.u., the AVD is decreased by 33.33%; when the target voltage was set at 1.00 p.u., the AVD is decreased by 36.54%. Due to the downsizing of DNN, the solving time is shortened by 12.12%.

Table VI and Fig. 11 show the statistical dispersion of the voltage. For the target voltage at 1.03 p.u. and 1.00 p.u., the WTs' average voltage with control policy train by the proposed method is 1.0300 p.u. and 1.0000 p.u. respectively; the voltage's standard deviation is 4.99×10^{-4} p.u. and 5.12×10^{-4} p.u. respectively, which shows that the proposed MADDPG-based method has better voltage control capacity.

Table V. Voltage deviation of each method

Control Policy Number	Target Voltage (p.u.)	AVD (×10-4 p.u.)	Solving Time (ms)	Algorithm
No Control	/	/	/	/
1	1.03	5.4	3.3	DDPG
2	1.03	3.6	2.9	MADDPG
3	1.00	5.2	/	DDPG
4	1.00	3.3	/	MADDPG

Table VI. Voltage statistical dispersion of each method

Control Policy Number	Target Voltage (p.u.)	Average Voltage (p.u.)	Standard Deviation (×10-4 p.u.)	Algorithm
No Control	/	1.0767	26.9	/
1	1.03	1.0299	6.80	DDPG
2	1.03	1.0300	4.99	MADDPG
3	1.00	1.0001	21.45	DDPG
4	1.00	1.0000	5.12	MADDPG

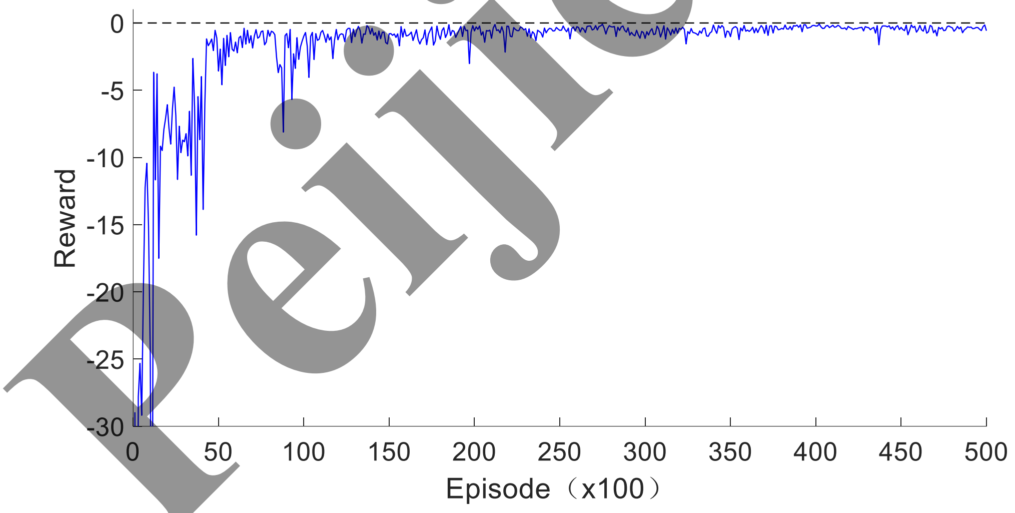

Fig. 5. The evolution of the reward during the training procedure.

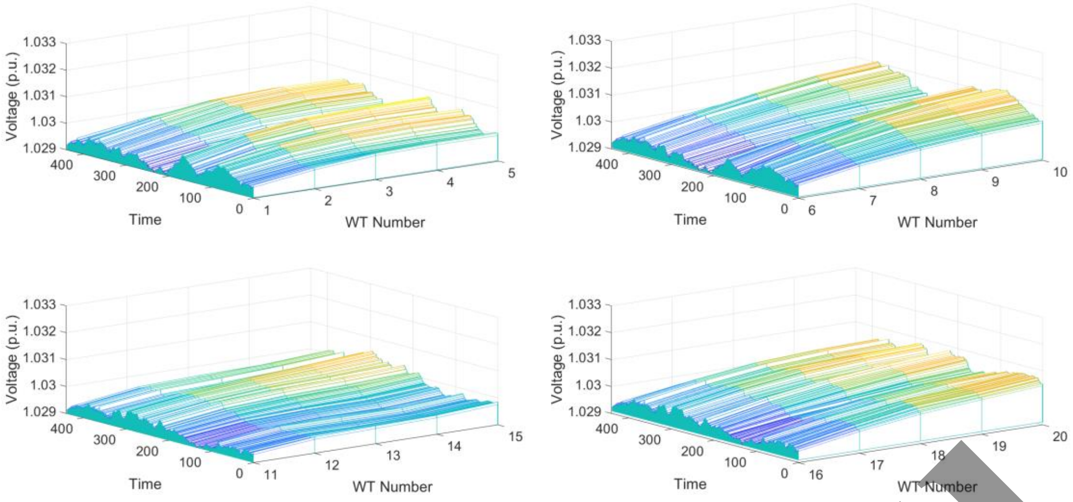

Fig. 6. The voltage distribution of the wind farm。

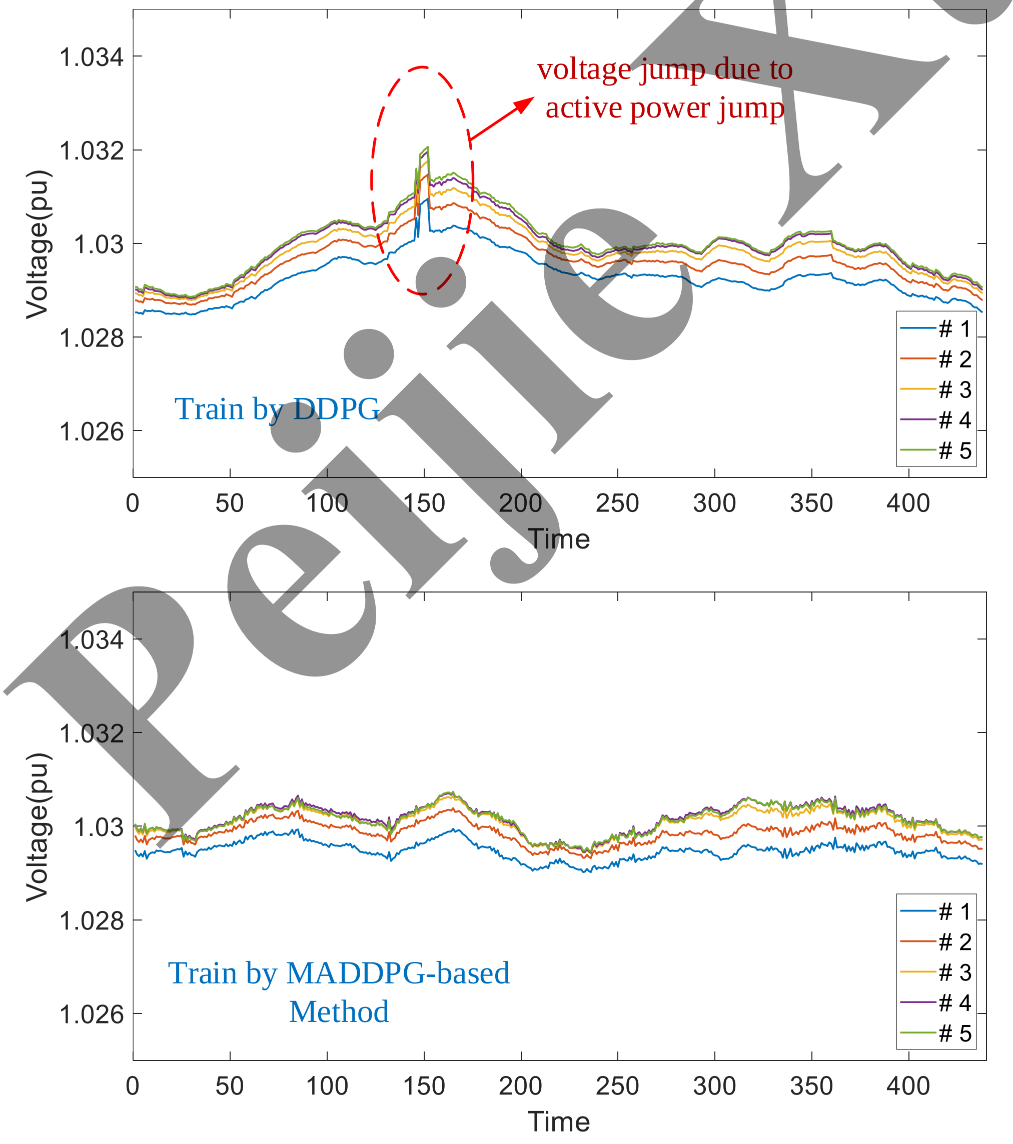

Fig. 7. Simulation results for voltage of WTs on the first feeder with aim at 1.03 p.u., using DDPG and proposed MADDPG-based method respectively.

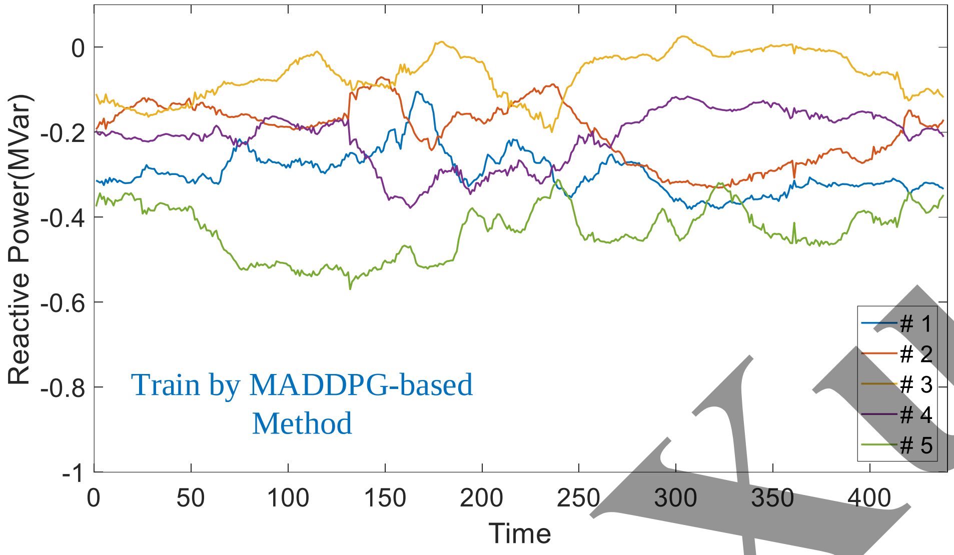

Fig. 8. Reactive power of WTs on a feeder, namely the action of the agent

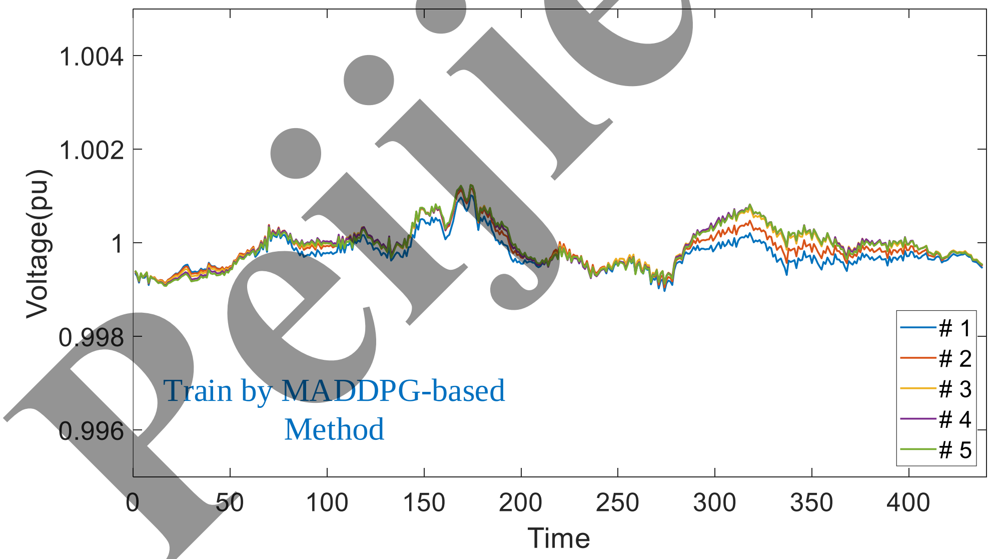

Fig. 9. Simulation results for voltage of WTs on the first feeder with aim at 1.00 p.u., using the proposed MADDPG-based method.

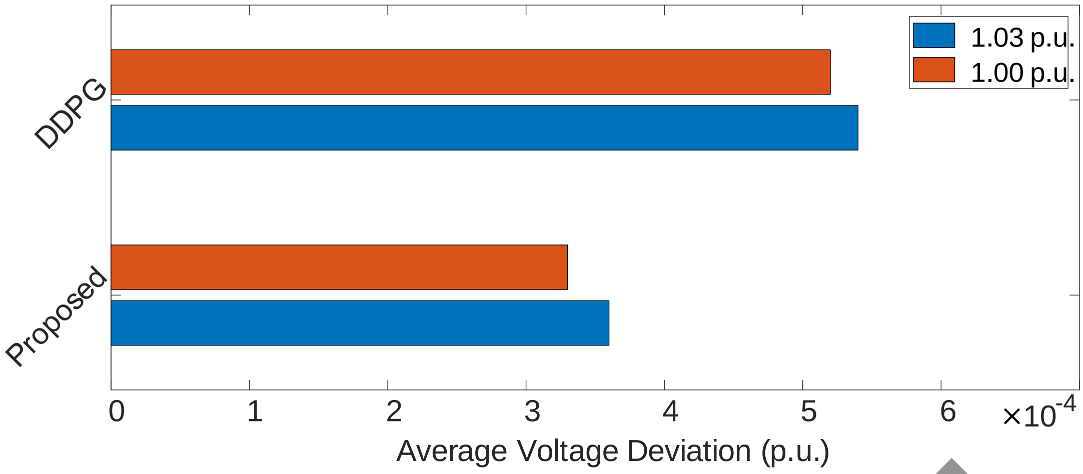

Fig. 10. Comparison of DDPG method and proposed MADDPG-based method, the less AVD means better control policy.

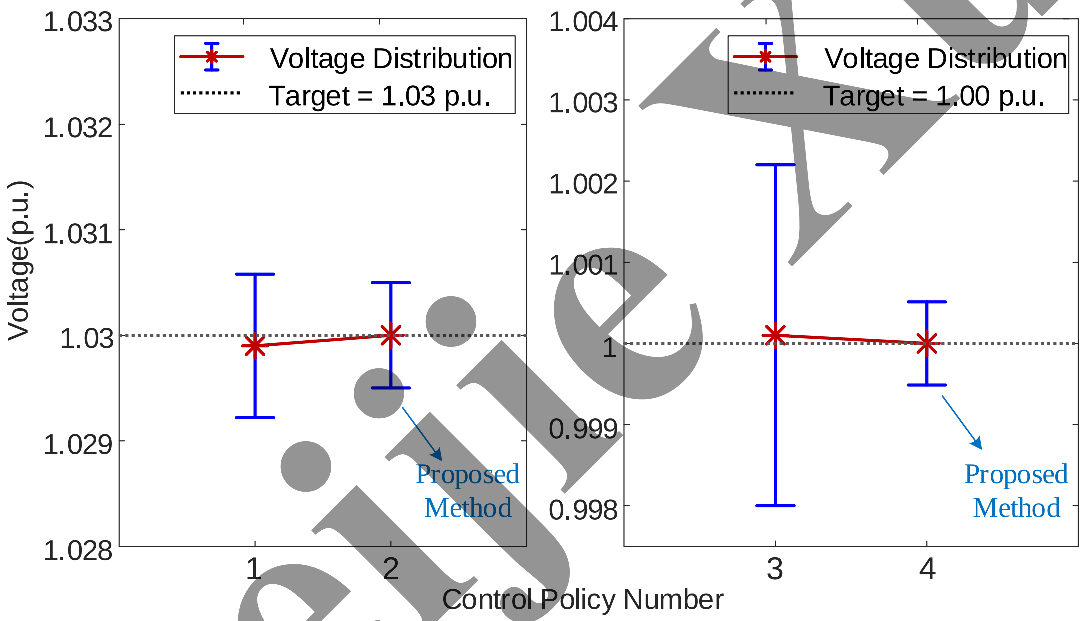

Fig. 11.

V Conclusion

A novel control architecture is proposed in this paper for distributed voltage control in an offshore wind farm. Trained by proposed MADDPG-based algorithm, the policy network efficiently control each WTs' voltage near the reference voltage using Q-V coordinated control. With the target of minimizing voltage deviation and apparent power overflow, each agent/feeder controls its WTs' reactive power only based on local observation, i.e. voltage, active power, and its forward difference. Therefore, the method significantly reduces the requirements of communications and knowledge of system parameters. In the future work, more consideration could be taken in, e.g. power loss between to nodes; furthermore, the reference voltage could also be add into observation.