Title

Model-free voltage control of active distribution system with PVs using surrogate model-based deep reinforcement learning

Authors

Cao, Di

Zhao, Junbo

Hu, Weihao

et al.

Publication Date

2022

DOI

10.1016/j.apenergy.2021.117982

Applied Energy 306 (2022) 117982

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Model-free voltage control of active distribution system with PVs using surrogate model-based deep reinforcement learning

Di Cao [a], Junbo Zhao [b], Weihao Hu [a,*], Fei Ding [c], Nanpeng Yu [d], Qi Huang [a], Zhe Chen [e]

[a] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
[b] Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762, United States
[c] Power Systems Engineering Center, National Renewable Energy Laboratory, Golden, United States
[d] Department of Electrical and Computer Engineering, University of California, Riverside, United States
[e] Department of Energy Technology, Aalborg University, Aalborg, Denmark

HIGHLIGHTS

- It proposes a physical-model-free control method of distribution network.
- It develops a method to handle unbalanced distribution system.
- It proposes a real-time control strategy to deal with fast voltage fluctuations.

ARTICLE INFO

Keywords:
Voltage regulation
Active distribution network
Model-free
Deep reinforcement learning
Solar PVs
Optimization

ABSTRACT

Accurate knowledge of the distribution system topology and parameters is required to achieve good voltage control performance, but this is difficult to obtain in practice. This paper proposes a physical-model-free voltage control method based on a surrogate-model-enabled deep reinforcement learning approach. Specifically, a surrogate model is trained in a supervised manner using the recorded limited number of historical data to learn the relationship between the power injections and voltage fluctuations of each node. Then, the deep reinforcement learning algorithm is applied to learn an optimal control strategy from the experiences obtained by continuous interactions with the surrogate model. The proposed method can achieve physical-model-free control of unbalanced distribution network and inform real-time decisions to deal with fast voltage fluctuations caused by the rapid variation of PV generation. Simulation results on an unbalance IEEE 123-bus system show that the proposed method can achieve similar performance as that of perfect physical-model-based approaches while being advantageous over other traditional methods.

1. Introduction

Active distribution network (ADN) is an effective way to improve the integration of distributed energy resources (DERs) since it can realize the local consumption and avoid unnecessary power loss from remote centralized generation resources [1]. However, the uncertainties and intermittence of DERs bring numerous technical challenges for ADN operations [2]. The penetration of renewable energy sources may change the voltage profiles and make the voltage cross the limits [3].

To better regulate the ADN voltage, various approaches have been proposed. For utilities, slow-acting devices, such as on-load tap changer (OLTC) [4], capacitor banks [5], and network reconfiguration [6], are widely used. But they could not deal with fast voltage fluctuations caused by PVs and demand responses. This motivates the curtailment of PV generations [7], reactive power control of PV inverter [8], and energy management of energy storage systems (ESSs) [9–10]. The curtailment of PV output reduces the economic benefits of customers and it cannot provide voltage support during the night. The integration of ESSs can enhance the reliability and flexibility of the system through balancing generation and demand [11], but it suffers from high investment and maintenance costs, and this has not been largely deployed in today's ADN. By contrast, reactive power control of PV inverters is an economically attractive solution since it does not cause the waste of solar power with negligible extra investments. Tests in [12] demonstrate

* Corresponding author.
E-mail address: whu@uestc.edu.cn (W. Hu).

https://doi.org/10.1016/j.apenergy.2021.117982
Received 3 August 2021; Received in revised form 17 September 2021; Accepted 29 September 2021
Available online 6 November 2021

that the reactive power control of PV inverters can achieve the most optimum economy as compared to the active power curtailment of PV, distributed ESS, and OLTC controls.

To deal with the uncertainties of load demand and DERs, stochastic programming (SP) and robust optimization (RO) are developed. A two-timescale SP based control method is proposed for the optimization of distribution network with renewable generators [13]. Ref. [14] proposes a SP-based reactive power scheduling strategy for the optimization of microgrids. SP methods require pre-sampling scenarios according to the assumed probability distribution. However, that information is difficult to obtain in practice. They also suffer from a heavy computation burden. RO achieves robust operation by constructing a solution that immunizes all possible realizations in the uncertainty set [15]. A two-stage control framework is proposed in [16] for the scheduling of PV inverters in ADN. Ref. [17] proposes a distributed control method for the optimization of distribution network based on network partition and distributed RO method. A RO-based approach is developed in [18] for the coordinated scheduling of active and reactive power in ADN. Both SP and RO-based control strategies are model-based. A common assumption is that the parameters and topology [19–20] of ADN are accurate, which is challenging to guarantee [21–22]. Moreover, SP and RO deal with uncertainties of DERs and loads by finding a predetermined solution. However, DERs can fluctuate a lot in a short time. For example, the PV output may change rapidly in a few seconds due to cloud dynamics [23]. In this condition, more frequent operations of controllable devices are required to cope with the fast-changing outputs of PVs. But they have to re-compute the optimal solutions and therefore are difficult to be used for real-time decisions.

To address the above issue, machine learning (ML)-based methods are developed. They can extract powerful operation knowledge from historical data to deal with the uncertain environment. Ref. [24] presents a data-driven local control method for the optimization of ADNs by mimicking the optimal behaviors achieved by the chance-constrained OPF method. By consolidating the OPF and the learning stage, [25] proposes a kernel-based approach for the control of smart inverters by explicitly considering the practical limitations. The forecasting information is first utilized to obtain the optimal inverter rules for the upcoming time intervals, which are then assigned to each inverter. During the execution stage, real-time decisions can be achieved by each inverter according to the derived rules and the collected inputs. Since the control rules are trained periodically, it can achieve a better compromise between computational complexity and communication requirements. However, all the aforementioned methods require accurate knowledge of the parameters and topology of the ADN, which is difficult to obtain in practice.

Various model-free approaches have been proposed in recent years. In [26], a model-free control method of VAR resources in balanced ADN using an extreme seeking (ES) algorithm is proposed. This method directly interacts with the real distribution network and thus is risky and costly. The deep reinforcement learning (DRL) method can implement offline training by interaction with the simulation model of ADN and then apply the trained agent to the real system [27–30]. A multi-time scale voltage control strategy is proposed in [27] combining DRL and physics-based optimization. Ref. [28] develops a double deep Q-learning-based method for the management of ESSs in a micro-grid. Ref. [29] proposes a multi-agent DRL-based approach for the coordinate control of PV inverters. Ref. [30] proposes a distributed control method based on network partition and multi-agent DRL algorithm. The aforementioned studies [27–30] rely on the model of ADN to calculate the reward during the training procedure, and thus the assumption on accurate knowledge of the parameters and topology of the ADN is still there [31]. Ref. [32] develops a physical model-free approach for the dynamic configuration of ADN based on augmented DRL. Synthetic operation data are generated to augment the original data set for training the DRL agent. Since there is no interaction between the agent and physical model of ADN during training, this method can achieve physical model-free control. However, it requires a large amount of training data and distribution mismatch may degrade the performance of the algorithm even when sufficiently large and diverse data are given [33]. Ref. [34] proposes a dynamic reconfiguration method of ADN based on a batch constrained DRL algorithm. However, many expert-level historical operational datasets are required for the training of the offline DRL algorithm, which is difficult to obtain in a practical system.

To bridge these gaps, this paper proposes a physical model-free approach for the voltage regulation of three-phase unbalanced ADN utilizing the reactive power capability of PV inverters and static var compensator (SVC). The main contributions are summarized as follows:

(1) The proposed approach synergistically integrates the deep neural network (DNN) based surrogate model with the DRL algorithm to achieve physical-model-free control. This is different from existing DRL approaches that need an extensive number of historical data or a good physical model.
(2) The proposed approach can handle an unbalanced three-phase distribution system while controlling single-phase PV inverters, SVCs, and active power curtailment of PV.
(3) The proposed method can inform decisions based on the latest observations in real-time to deal with fast voltage fluctuations caused by the rapid variation of PV generations.

The rest of this paper is organized as follows. Section 2 describes the mathematical model of the voltage regulation problem. In Section 3, the surrogate model and control are illustrated in detail. Numerical results are discussed in Section 4. Section 5 concludes the paper.

2. Problem statement

2.1. System model and constraints

Consider an ADN of $N+1$ buses served by the substation indexed by $n=0$. All the buses of the ADN are collected into $N_0 := \{0\} \cup N$. For each bus $i \in N$, let v_i^φ represents its complex voltage of phase $\varphi \in \{a,b,c\}$, and $p_i^\varphi + jq_i^\varphi$ the injected complex power of phase φ. The active power injection p_i^φ is split into $p_i^\varphi := p_{i,g}^\varphi - p_{i,c}^\varphi$, where $p_{i,g}^\varphi$ and $p_{i,c}^\varphi$ denote the active power generation and consumption, respectively. Likewise, the reactive power injection can be decomposed into $q_i^\varphi := q_{i,g}^\varphi - q_{i,c}^\varphi$, where $q_{i,g}^\varphi$ and $q_{i,c}^\varphi$ represent the reactive power generation and consumption, respectively. The voltage is decomposed into $v_i^\varphi := e_i^\varphi + jf_i^\varphi$, where e_i^φ and f_i^φ denote the real and imaginary parts of the complex voltage of phase φ at bus i, respectively. If only load demand is connected to phase φ at bus i, then it holds that $p_{i,g}^\varphi = q_{i,g}^\varphi = 0$. When phase φ at bus i is equipped with a distributed generator, then $p_{i,c}^\varphi \geqslant 0,\ q_{i,c}^\varphi \geqslant 0,\ p_{i,g}^\varphi \geqslant 0$.

The active and reactive power flow constraints are expressed as

$$p_{i,g}^\varphi - p_{i,c}^\varphi - e_i^\varphi \sum_{j=1}^{N} \sum_{\alpha=a,b,c} (G_{ij}^{\varphi\alpha} e_j^\alpha - B_{ij}^{\varphi\alpha} f_j^\alpha) - f_{i,t}^\varphi \sum_{j=1}^{N} \sum_{\alpha=a,b,c} (G_{ij}^{\varphi\alpha} f_j^\alpha + B_{ij}^{\varphi\alpha} e_j^\alpha) = 0, i \in N \tag{1}$$

$$q_{i,g}^\varphi - q_{i,c}^\varphi - f_i^\varphi \sum_{j=1}^{N} \sum_{\alpha=a,b,c} (G_{ij}^{\varphi\alpha} e_{j,t}^\alpha - B_{ij}^{\varphi\alpha} f_{j,t}^\alpha) + e_i^\varphi \sum_{j=1}^{N} \sum_{\alpha=a,b,c} (G_{ij}^{\varphi\alpha} f_{j,t}^\alpha + B_{ij}^{\varphi\alpha} e_{j,t}^\alpha) = 0, i \in N \tag{2}$$

where $G_{ij}^{\varphi\alpha}$ and $B_{ij}^{\varphi\alpha}$ represent the real and imaginary parts of the complex admittance matrix elements. The relationships between the three phases of root bus s are expressed as

$$\begin{cases} f_s^a - e_s^a tan(\frac{0\pi}{180}) = 0 \\ f_s^b - e_s^b tan(\frac{-120\pi}{180}) = 0 \\ f_s^c - e_s^c tan(\frac{120\pi}{180}) = 0 \end{cases} \tag{3}$$

The overvoltage will trigger the protection device to cut off the distributed generator. Therefore, the voltage constraint is considered

$$V_{min}^2 \leqslant (e_i^\varphi)^2 + (f_i^\varphi)^2 \leqslant V_{max}^2, i \in N \tag{4}$$

where V_{min} and V_{max} are the lower and upper bounds of voltage. This study follows the ANSI-C.84.1 standard, which specifies that the voltage deviation of each node should be within ±5% p.u.

2.2. Controllable device models and constraints

The controllable devices in ADN can be divided into two types: mechanical devices and power electronic devices. Mechanical devices include the on-load tap changer, voltage regulator, and switched capacitor. Since the limited life cycle and slow response speed of mechanical devices, they are typically scheduled offline, making it difficult for them to deal with frequent voltage fluctuation caused by the growing deployment of renewable energy generators. By contrast, SVC is a power electronic device, which can provide reactive power support within seconds in a continuously valued fashion. This makes it a promising solution to engage SVC in voltage regulation of ADN with a high-level penetration of renewable energy. In addition, PV units are typically equipped with smart inverters. Ref. [35] demonstrates that the utilization of smart inverters can enhance power quality and reduce energy loss. The revised standards by IEEE 1547.8 working group allow the smart inverters to provide reactive power support for the voltage regulation of ADN [36]. Considering the above benefits, PV inverters are used for voltage control in this study. Suppose that the total number of SVC and PV inverters is M. We collect all the buses with controllable devices into G. For each bus $j \in G$, let $q_{j,SVC}^\varphi$ represent the reactive power generated by SVC that is connected to phase φ, and $p_{j,PV}^\varphi / q_{j,PV}^\varphi$ the active/reactive power generation of PV connected to phase φ. The generated reactive power is decomposed into $q_{j,g}^\varphi := q_{j,SVC}^\varphi + q_{j,PV}^\varphi$. Since only PVs can generate active power in the studied system, it holds that $p_{j,g}^\varphi = p_{j,PV}^\varphi$. The reactive power generated by SVC is constrained as

$$q_{SVC,min} \leqslant q_{j,SVC}^\varphi \leqslant q_{SVC,max}, j \in G \tag{5}$$

where $q_{SVC,min}$ and $q_{SVC,max}$ represent the lower and upper limits of the reactive power generated by SVC. PV curtailment is an effective way to deal with the overvoltage of ADN. The curtailment of PV generation $p_{j,cur}^\varphi$ is constrained by

$$0 \leqslant p_{j,cur}^\varphi \leqslant \beta p_{j,PV}^\varphi, j \in G \tag{6}$$

where β represents the maximum curtailment ratio of PV generator. To improve the utilization ratio of renewable energy, the PV curtailment is not allowed to cross a certain level. Then the active power injection $p_{j,g}^\varphi$ is decomposed into $p_{j,g}^\varphi := p_{j,PV}^\varphi - p_{j,cur}^\varphi$. The reactive power output of the PV inverter is constrained as

$$(p_{j,PV}^\varphi)^2 + (q_{j,PV}^\varphi)^2 \leqslant (s_{j,PV}^\varphi)^2, j \in G \tag{7}$$

where $s_{j,PV}^\varphi$ represents the apparent power of the PV inverter connected to phase φ at bus j. Typically, the apparent power of the PV inverter is set to 1.0–1.1 times the rated capacity of the PV unit. The oversized inverter can provide reactive support for voltage regulation even when the maximum output of the PV generator is reached. Choosing $s_{j,PV}^\varphi = 1.08 p_{j,PV}^\varphi$, for example, the maximum reactive power that can be generated by the inverter is about 40% of rated active power capability when the rated capacity of PV unit is reached, demonstrating that the reactive power support ability of the inverter can be improved by increasing the apparent power of the inverter.

2.3. Voltage regulation formulation

Given load demand, PV generation, and parameters of the ADN, the goal of the voltage regulation problem is to decide $q_{j,PV}^\varphi$, $q_{j,SVC}^\varphi$ and $p_{j,cur}^\varphi$ to minimize the voltage deviations and the amount of PV curtailment while satisfying corresponding constraints

$$\min_{q_{j,PV}^\varphi, q_{j,SVC}^\varphi, p_{j,cur}^\varphi} F(x) = \sum_{\varphi=a,b,c} \sum_{i=1}^{N} \left| \sqrt{(e_i^\varphi)^2 + (f_i^\varphi)^2} - V_0 \right| + \delta \sum_{j=1}^{G} p_{j,cur}^\varphi \quad \text{Subject to } (1)-(7) \tag{8}$$

Solving this optimization problem typically requires the accurate parameters of the ADN, which are difficult to obtain in practice and affected by many uncertainties. In addition, it is difficult for traditional methods to deal with fast voltage fluctuations caused by the rapid variation of PV generations. To this end, a physical-model-free real-time control strategy based on surrogate-model-enabled DRL method is proposed in this study.

3. Proposed model-free voltage control framework

The proposed model-free control framework is shown in Fig. 1. It consists of three main parts, namely the MDP formulation, DRL control, and building of surrogate, actor, and critic networks.

3.1. MDP formulation

In this paper, the voltage regulation problem is formulated as an MDP with finite time-steps and we mainly focus on the design of the following components:

- **Agent**: the agent refers to the controller that is in charge of the control of PV and SVC.
- **State-space S**: the state at time slot t, $s_t \in S$ consists of three components: $(p_{i,c}^\varphi, p_{j,PV}^\varphi, q_{i,c}^\varphi), i \in N, j \in G$.
- **Action space A**: the action at time slot t, $a_t \in A$ consists of three components: $(q_{j,PV}^\varphi, q_{j,SVC}^\varphi, p_{j,cur}^\varphi)$, $j \in G$, see (5)-(7) for detailed explanations about the variables.
- **Reward function R**: the objective of the model is to reduce the voltage deviations and PV curtailments. Thus, the immediate reward of an agent at time slot t is $r_t = -\left(\sum_{i=1}^{N} \sum_{\varphi=a,b,c} \left| \sqrt{(e_i^\varphi)^2 + (f_i^\varphi)^2} - V_0 \right| + \delta \sum_{j=1}^{G} p_{j,cur}^\varphi \right) - \eta$, where η is the penalty term when the voltage crosses the threshold and will be further elaborated in Section 4.1.

One MDP is composed of a finite number of time steps. At each time slot, the agent decides the control action a_t based on the observed state s_t, obtains an immediate reward r_t. The objective of the agent is to learn a deterministic voltage regulation strategy $a_t = \pi(s_t)$ to maximize the discounted cumulative reward from the current time step onward $R(s_t, a_t) = r_t + \gamma r_{t+1} + \dots + \gamma^{T-t} r_T$, where $\gamma \in [0, 1]$ is the discount factor to balance the future reward against the immediate reward [37].

3.2. DRL control model

DDPG is an actor-critic framework-based algorithm, which simultaneously optimizes two functions for the problem. The policy function maps the state s_t to the desired output a_t. The critic function maps state

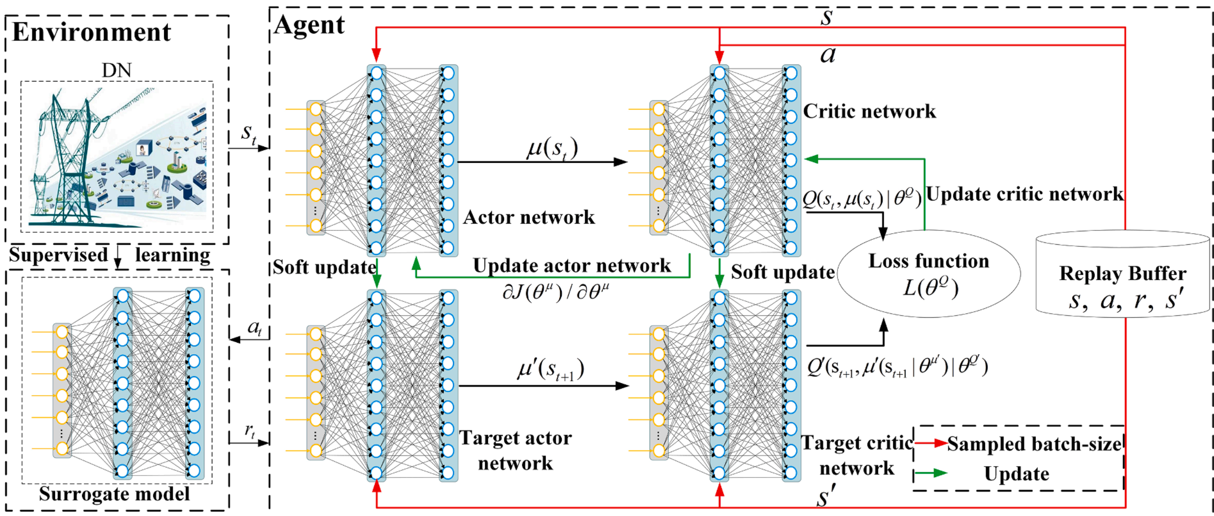

Fig. 1. Proposed model-free control framework integrating surrogate modeling and DRL.

and action pairs (s_t, a_t) to the expected cumulative reward. They are trained against each other such that the critic function better predicts the outcomes, and the policy function produces control decisions with reduced voltage deviations, see Fig. 1 for the details.

(1) **Critic function**

The critic function is also named the action-value function $Q(s_t, a_t)$, which is the expected cumulative reward when action a_t is taken in the state s_t under the policy π. The parameters of the action-value function θ^Q can be optimized by minimizing the following loss function [38]:

$$L(\theta^Q) = \mathbb{E}_{\mu'}[(Q(s_t, a_t|\theta^Q) - y_t)^2] \tag{9}$$

$$y_t = r(s_t, a_t) + \gamma Q(s_{t+1}, u(s_{t+1})|\theta^Q) \tag{10}$$

The critic network is trained in a supervised learning manner and y_t and $Q(s_t, a_t|\theta^Q)$ should be as close as possible.

(2) **Policy function**

The policy function maps the state to action. Ref. [38] shows that the parameters of the policy function θ^μ should be updated towards the gradient of $J(\theta^\mu)$, i.e.,

$$\frac{\partial J(\theta^\mu)}{\partial \theta^\mu} = \mathbb{E}_{s\sim\rho^\mu}[\nabla_{\theta^\mu} Q(s, a|\theta^Q)|_{s=s_t, u=u(s_t|\theta^\mu)}] = \mathbb{E}_{s\sim\rho^\mu}[\nabla_\theta \mu_\theta(s|\theta^\mu)|_{s=s_t} \nabla_a Q(s, a|\theta^Q)|_{s=s_t, u=u(s_t|\theta^\mu)}] \tag{11}$$

The parameters of the policy network are adjusted in the direction that maximizes the Q value, which is the expected cumulative reward the agent achieves at a state.

(3) **Target networks**

To enhance stability during training, DDPG introduces a target critic function $Q'(\cdot)$ and a target actor function $\mu'(\cdot)$ for the calculation of the targets y_t. Equation (10) is thus rewritten as [38]

$$y_t = r(s_t, a_t) + \gamma Q'(s_{t+1}, \mu'(s_{t+1}|\theta^{\mu'})|\theta^{Q'}) \tag{12}$$

The parameters of the target actor function $\theta^{\mu'}$ and critic function $\theta^{Q'}$ are updated by slowly tracking the online neural networks: $\theta^{\mu'} \leftarrow \tau\theta + (1-\tau)\theta^\mu$, $\theta^{Q'} \leftarrow \tau\theta + (1-\tau)\theta^Q$, where $\tau \ll 1$.

(4) **Replay buffer and exploration**

During the training process, the input data should be independent and identically distributed. For the DRL algorithm, the data are correlated with each other. To improve its stability, the experience replay mechanism is adopted. The data are continuously stored in the replay buffer, from which a batch size is uniformly sampled to train the DNN. This mechanism helps to reduce the correlation among the sequence data [38].

3.3. Building of surrogate, actor, and critic model

(1) **DNN for surrogate model**

Surrogate modeling aims to learn the nonlinear mapping from the active and reactive power injections to the voltage magnitude of each node. DNN is a powerful algorithm that has a strong nonlinear fitting ability [39] and feature extraction capability [40]. By transforming the raw input from layer to layer hierarchically, it can learn high-dimensional abstract feature representations from the training data [41]. The input of the surrogate model includes the active power injection p_i^φ and reactive power injection q_i^φ of each node. The outputs are the voltage magnitude of each node $|v_i^\varphi|$. We collect the nodal power injection and voltage magnitude into vector P, Q, and V, respectively. The relationship between the voltage and active and reactive power injections can be represented as:

$$V = \langle W^s, g^s(P, Q)\rangle + b^s \tag{13}$$

where $\langle \cdot, \cdot \rangle$ is the inner product; W^s is the weight matrix of the output layer; b^s is the bias of the output layer; $g^s(\cdot)$ denotes the hierarchical transformation of the input through multi-layer nonlinear mappings. In this paper, multiple fully connected hidden layers are used. Therefore, $g^s(\cdot)$ is

$$g^s(V) = tanh(W_L^s \cdot v_L^s(P, Q) + b_L^s) \tag{14}$$

$$v_{l+1}^s(P, Q) = tanh(W_l^s \cdot v_l^s(P, Q) + b_l^s),\ l = 1, \ldots, L+1 \tag{15}$$

$$v_1^s(P, Q) = (P, Q) \tag{16}$$

where W_l^s is the weight matrix of the lth layer; b_l^s is the bias of the lth layer; $tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$ is the activation function of hidden layers. The parameters of the surrogate model are represented as $\theta^s = \{W_1^s, b_1^s, \ldots, W_L^s, b_L^s, W^s, b^s\}$.

It is worth noting that using the same training data, the system parameters may be estimated and the voltage at each node can be calculated [42]. However, estimation of line parameters and topology of unbalanced distribution networks is difficult as it is affected by many factors. Since we aim to obtain the mapping relationships between voltage magnitudes and node power injections, it is much easier to learn. We train a surrogate model in a supervised manner to represent that mapping relationship and then interact with the DRL agent to learn optimal control policy.

(2) **DNN for critic function**

Due to the strong non-linear fitting ability of DNN, it is used to approximate the critic function. The inputs of the critic network include two components: the state s_t and action a_t. The state s_t includes the active power of each node $p_{i,c}^{\varphi}$, $i \in N$, the reactive power consumption of each node $q_{i,c}^{\varphi}$, $i \in N$, and the active power generation of each PV unit $p_{j,g}^{\varphi}$, $j \in G$. The action includes the reactive power generated by each PV inverter $q_{j,PV}^{\varphi}$, $j \in G$, the reactive power of each SVC $q_{j,SVC}^{\varphi}$, $j \in G$, and the PV curtailment $p_{j,cur}^{\varphi}$, $j \in G$. The output of the critic is $Q(s_t, a_t)$, which is a scalar representing the value of action a_t under the current state s_t. The relationship between the inputs and output can be expressed as

$$Q(s_t, a_t) = \langle W^Q, g^Q(s_t, a_t) \rangle + b^Q \tag{17}$$

where $\langle \cdot, \cdot \rangle$ denote the inner product operation; W^Q and b^Q represent the weight matrix and bias of the output layer of the critic network, respectively. $g^Q(s_t, a_t)$ denotes the latent features extracted by the multiple fully-connected hidden layers:

$$g^Q(s_t, a_t) = ReLU(W_L^Q \cdot v_L^Q(s_t, a_t) + b_L^Q) \tag{18}$$

$$v_{l+1}^Q(s_t, a_t) = ReLU(W_l^Q \cdot v_l^Q(s_t, a_t) + b_l^Q),\ l = 1, \ldots, L+1 \tag{19}$$

$$v_1^Q(s_t, a_t) = (s_t, a_t) \tag{20}$$

where W_l^Q and b_l^Q are the weight matrix and bias of the lth layer of critic network, respectively; $ReLU(x) = max(0, x)$ is the specified activation function. Then the parameters of the critic network are $\theta^Q = \left\{W_1^Q, b_1^Q, \ldots, W_L^Q, b_L^Q, W^Q, b^Q\right\}$.

(3) **DNN for Actor Function**

To deal with the dynamic environment, this paper use DNN to approximate the policy function. The inputs of the actor-network are the state of the MDP s_t, which includes the active power consumption of each node $p_{i,c}^{\varphi}$, $i \in N$, the reactive power consumption of each node $q_{i,c}^{\varphi}$, $i \in N$, and the active power generation of each PV unit $p_{j,g}^{\varphi}$, $j \in G$. The outputs of the actor-network are the action of the MDP a_t, which is composed of the reactive power of the PV inverter $q_{j,PV}^{\varphi}$, $j \in G$, the reactive power of SVC $q_{j,SVC}^{\varphi}$, $j \in G$, and the PV curtailment $p_{j,cur}^{\varphi}$, $j \in G$. DNN is utilized to approximate the actor function via

$$a_t = tanh(\langle W^\mu, g^\mu(s_t) \rangle + b^\mu) \tag{21}$$

where $\langle \cdot, \cdot \rangle$ is the inner product; W^μ is the weight matrix of the output layer of actor-network; b^μ represents the bias of the output layer of actor-network; $tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$ is the activation function of the output layer, the range of which is $(-1,\ 1)$; $g^\mu(s_t)$ denotes the latent feature extracted from the input state s_t by the multiple fully-connected hidden layers of the actor-network, which can be derived by

$$g^\mu(s_t) = ReLU(W_L^\mu \cdot v_L^\mu(s_t) + b_L^\mu) \tag{22}$$

$$v_{l+1}^\mu(s_t) = ReLU(W_l^\mu \cdot v_l^\mu(s_t) + b_l^\mu),\ l = 1, \ldots, L+1 \tag{23}$$

$$v_1^\mu(s_t) = s_t \tag{24}$$

where W_l^μ is the weight matrix of the lth layer of actor-network; b_l^μ denotes the bias of the lth layer of actor-network; $ReLU(\cdot)$ denotes the rectified linear units function, which is the activation function. The parameters to be optimized of actor-network are collected into $\theta^\mu = \{W_1^\mu, b_1^\mu, \ldots, W_L^\mu, b_L^\mu, W^\mu, b^\mu\}$.

3.4. Training process of the proposed method

There are three sets of parameters to be optimized: parameters of the surrogate model θ^s, critic network θ^Q and policy network θ^μ. The training process can be divided into two steps, which are shown in Table 1. In the first step, the surrogate model is trained in a supervised manner. At each epoch, batches of instances are sampled to calculate the loss according to the mean square error:

$$L(\theta^s) = \frac{1}{B} \sum_{i=1}^{B} \left[\left(V_i - \widehat{V}_i(P_i, Q_i | \theta^s)\right)^2 \right] \tag{25}$$

where $\widehat{V}_i(\cdot)$ is the predicted value of voltage. Then, stochastic gradient descent is applied to update the parameters θ^s via

Table 1
Training of the proposed method.

Algorithm Training of the proposed method

1: Randomly initialize the parameters of the surrogate model θ^s
2: For epoch $= 1, 2, \ldots, M$ do
3: Sample batch from the training set $\{P_k, Q_k, V_k\}_{k=1}^{B}$
4: Optimize θ^s according to equations (25) and (26)
5: End for
6: Fix surrogate model parameters θ^s
7: Randomly initialize critic network $Q(s, a | \theta^Q)$ and actor-network $u(s | \theta^\mu)$ with weights θ^Q and θ^μ
8: Initialize target network Q' and u' with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$
9: For episode $= 1, 2, \ldots, N$ do
Receive initial observation s_1
For $t = 1, 2, \ldots 24$ do
10: Choose action a_t, execute the action and transfer to the next state s_{t+1}
11: calculate reward r_t based on surrogate model
12: Store transition (s_t, a_t, r_t, s_{t+1}) in the replay buffer
13: If the replay buffer is full: $\sigma \leftarrow \sigma^* \xi$
14: Sample a random mini-batch of transitions from the replay buffer
15: Update the critic-network via equations (28)-(29)
16: Update the actor-network via equations (30)-(31)
17: Update the target actor and critic networks through $\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$, $\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$
18: End for
19: End for

$$\theta_{t+1}^{s} = \theta_{t}^{s} - \lambda_s \nabla_{\theta^s} L(\theta^s) \tag{26}$$

where λ_s is the learning rate of the surrogate model. When training is completed, θ^s is fixed and the surrogate model is embedded into the environment of the DDPG algorithm.

In the second step, the parameters of the DDPG method are optimized. Specifically, the parameters are initialized randomly and the parameters of the target networks are copied from the online network. Then, the algorithm runs for N episodes to learn the voltage regulation strategy. One epoch corresponds to a randomly sampled day from the training set. Each epoch is divided into 24 time-steps, each corresponds to an hour in the day. For each time step, the agent obtains an observation of the environment s_t, chooses an action a_t, then calculates the reward r_t based on the surrogate model and the environment transfers to the next state s_{t+1}. The transition (s_t, a_t, r_t, s_{t+1}) is then stored in the memory buffer. The actions are chosen according to an exploration strategy with Gaussian noise:

$$a_t = \mu(s_t|\theta_t^{\mu}) + N(\mu(s_t|\theta_t^{\mu}), \sigma) \tag{27}$$

In the beginning, σ is a constant. When the memory capacity reaches the upper limit, σ attenuates at a fixed rate. At the same time, n batches of experiences $(s_j, a_j, r_j, s_{j+1}), j = 1, 2, \dots, n$ are randomly sampled from the memory to update θ^Q and θ^{μ}. In particular, θ^Q is updated by minimizing the following loss:

$$L(\theta^Q) = \frac{1}{N}\sum_{i=1}^{N}[(Q(s_t, a_t|\theta^Q) - y_t)^2] \tag{28}$$

θ^Q is optimized by the gradient descent:

$$\theta_{t+1}^{Q} = \theta_{t}^{Q} - \lambda_Q \nabla_{\theta^Q} L(\theta^Q) \tag{29}$$

where λ_Q is the learning rate of the critic network. θ^{μ} is updated according to the policy gradient:

$$\nabla_{\mu} J(\mu) = \frac{1}{N}\sum_{i=1}^{N}[\nabla_{\theta^{\mu}}\mu(s|\theta^{\mu})|_{s=s_i}\nabla_a Q(s, a|\theta^Q)|_{s=s_i, u=u(s_i|\theta^{\mu})}] \tag{30}$$

$$\theta_{t+1}^{\mu} = \theta_{t}^{\mu} - \lambda_{\mu} \nabla_{\theta^{\mu}} L(\theta^{\mu}) \tag{31}$$

where λ_{μ} is the learning rate of the policy network. After that, the parameters in the target networks are updated by slowly tracking θ^Q and θ^{μ}. When the training process is completed, the parameters of the actor-network are used for control.

4. Numerical results

Simulations are carried out on an unbalanced IEEE 123-bus system to evaluate the performance of the proposed method. Comparative results with various benchmark methods are also provided to illustrate the advantages of the proposed method.

4.1. Experimental setup

The schematic of an IEEE 123-bus system is shown in Fig. 2 [43]. To simulate high PV penetration, 9 PVs are connected to the ADN. The specifications of the installed PVs and SVCs are listed in Table 2. The maximum voltage deviation is set as ±5% of its nominal value, yielding the upper and lower bounds as 1.05 p.u. and 0.95 p.u., respectively. η is set to −20 in this study.

The proposed method has a surrogate model and a control model. The surrogate model is trained in a supervised manner to learn the

Table 2
Specifications of the controllable devices.

Type	Capacity	Locations
PV	0.6 MW/0.66MVA	9, 27, 43, 62, 75, 83, 91, 101, 112
SVC	0.3MVar	11, 50, 79

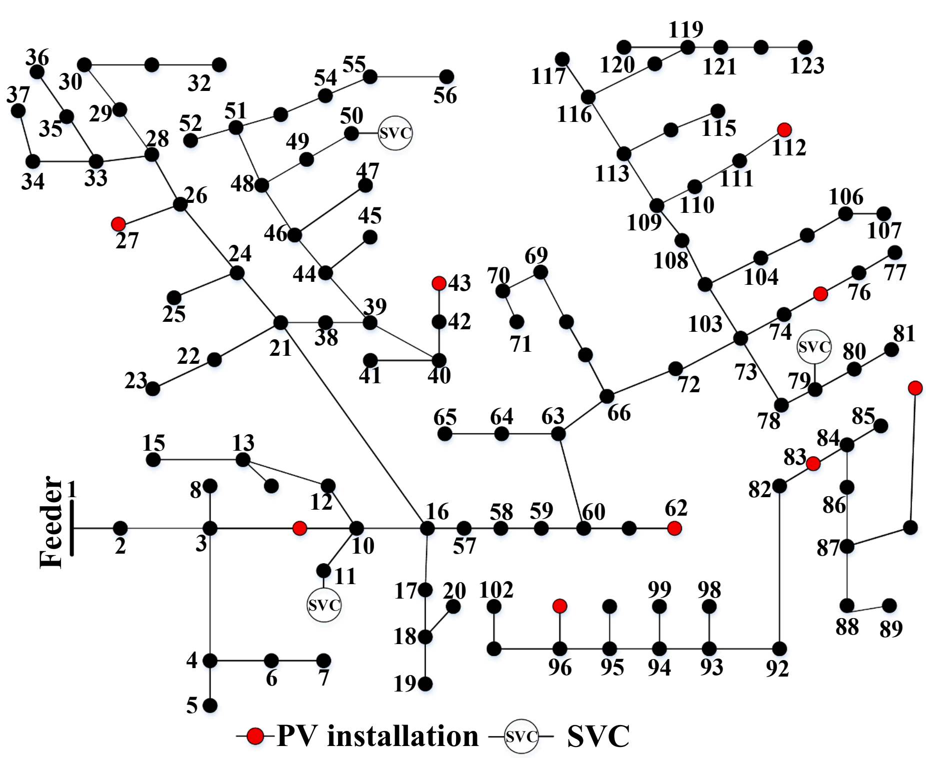

Fig. 2. Schematic of the IEEE 123-bus test system.

mapping between the voltage and the active and reactive power of the nodes. 4000 instances of $\{p_i^{\varphi}, q_i^{\varphi}, v_i^{\varphi}\}_{i=1,2,\dots,N\ \varphi=a,b,c}$ data are generated by the three-phase AC power flow model. In particular, 4000 instances load data are obtained, i.e., the load data are multiplications of three parts: the first part is the baseload for each node; the second part is the time-coefficient, which varies with the time index while the third part is the random coefficient for each node that is randomly sampled from [0.8, 1.2]. For the PV generation data, the field measurement data of Xiaojin, a county of Sichuan province of China are utilized. The data are first normalized according to the maximum PV generation of this county. Then the normalized data are scaled according to the PV capacity in the ADN. 4000 instances of PV generation data are randomly selected from the data set. After that, the load demands are randomly combined with the PV data to formulate 4000 instances of injected power $\{p_i^{\varphi}, q_i^{\varphi}\}_{i=1,2,\dots,N\ \varphi=a,b,c}$. Finally, the voltages at each instance $\{v_i^{\varphi}\}_{i=1,2,\dots,N\ \varphi=a,b,c}$ are calculated by the three-phase AC power flow model. The data are divided into two parts: a training set and a test set. The numbers of instances for the two parts are 3500 and 500, respectively. To avoid overfitting of the proposed method, the K-fold cross-validation method is applied to select the hyper-parameters of the DNN. Specifically, the training data is divided into 7 parts equally, one of which is used as the validation data and the rest as the training data. At every time, one part of the data is selected as the validation set to analyse the accuracy of the model trained using the rest part of the data. This process is repeated 7 times until all data have been utilized as the validation set. The hyper-parameters of the model that achieve the best performance during this process are utilized for the surrogate model. When the training process is finished, the test set is utilized to analyse the accuracy of the surrogate model. The parameter settings of the surrogate model are shown in Table 3.

For the training of the DRL agent, the PV generation data of Xiaojin are also used. The data are divided into two parts: the training and test sets, which contain 300 and 10 days' data, respectively. For the load data, 7440 instances of random coefficients for each node are randomly sampled between 0.8 and 1.2. The random coefficients are separated into two parts: the training and test sets, which contain 7200 and 240 instances of data, respectively. The control model is composed of actor-networks and critic networks, both of which share the same architecture. The parameter settings of the control model are shown in Table 4. The proposed method is written in Python with Keras. A workstation with an Intel Core i9-10980XE CPU is used for the training procedure.

4.2. Performance evaluation of the surrogate model

The surrogate model is trained for 5000 epochs to learn the complex mapping relationships. The objective is to minimize the quadratic loss function during training. The loss-epoch curve of the surrogate model during the training process is plotted in Fig. 3. Log scale is selected for better visualization. It can be observed from the figure that the loss is relatively high at the beginning of the training procedure. It gradually decreases during training and finally converges, indicating that the proposed DNN-based surrogate model gradually learns the mapping relationship between the voltage magnitude and power injection of each node.

To evaluate the performance of the surrogate model on the test set, the evolution of accuracy on test data during the training process is shown in Fig. 4. The mean absolute error (MAE) is used as the evaluation index, which is defined as follows:

$$MAE = \frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{i=1}^{N} |\widehat{v}_{m,i} - v_{m,i}| \tag{32}$$

where M is the number of total instances in the test set; N is the number of nodes in ADN; $\widehat{v}_{m,i}$ and $v_{m,i}$ are the predicted voltage by the learned surrogate model and the true voltage of node i at the mth instance in the test set, respectively. The MAE index represents the averaged absolute value of voltage prediction error at one node. The surrogate model is evaluated on test data every 50 epochs. At the beginning of the training, the MAE on test data is 0.013 p.u., which is relatively high. With the training process going forward, the MAE gradually decreases. When the training process is finished, the MAE on the test set is 1.3e−3 p.u., demonstrating that the prediction voltage is very close to the true voltage. The distribution of the prediction error of each node on the test set is shown in Fig. 5. It can be observed that the voltage prediction errors fall into a narrow range, demonstrating the effectiveness of the proposed surrogate model.

Table 3
Parameter setting of the surrogate model.

Parameter	Value
Number of neurons in hidden layers	400/200/200
Batch size for updating NN	32
Learning rate	0.0001
Number of training instances	3500
Maximum training epochs	5000

Table 4
Parameter setting of the drl model.

Parameter	Value
Neuron numbers of hidden layers	400/200
Batch size for updating NN	256
Step size of each episode	24
Learning rate for actor-network	0.001
Learning rate for critic-network	0.002
Discount factor	0.1
Maximum training episodes	50,000
Replay buffer size	100,000
Soft update coefficient	0.01

4.3. Performance evaluation of the control model

The training set is used to train the proposed control model. In this test, a comparative test is carried out among the proposed method and a physical-model-free method based augmented DRL [32], which trains DDPG agent utilizing batches of experience data (B-DDPG) sampled from the replay buffer. For the B-DDPG method, the experience data (s_t, a_t, r_t, s_{t+1}) in the replay buffer are constructed by the surrogate model based on historical data. At each time-step, a batch of experience data is sampled from the replay buffer for the training of the DDPG agent. Since there is no interaction between the DDPG agent and the ADN, the B-DDPG method is also a physical model-free approach. The performances of three B-DDPG models are evaluated when different numbers of experiences are stored in the replay buffer. The memory capacities of the B-DDPG1, B-DDPG2, and B-DDPG3 are 5000, 100000, 200000, respectively. The changes of the cumulative rewards obtained by various methods during the training process are shown in Fig. 6 with 50,000 episodes. The cumulative curves are averaged over 3 random seeds. It can be observed that when 5000 instances of experiences are stored in the memory, the B-DDPG method fails to learn a good control strategy. When we increase the number of experiences in the replay buffer, the cumulative reward obtained by the B-DDPG method improves. It achieves the highest cumulative reward when the memory capacity is set to 200000. As we continue to increase the capacity of the replay buffer, no improvement of cumulative reward is observed. By contrast, the cumulative reward of the proposed method increases significantly during the training process and finally converges around −55, which is higher than that obtained by B-DDPG3. The reason is that the experiences generated by continuous interaction and exploration can help the agent learn a better critic than offline synthetic experiences. The results demonstrate the advantages of the proposed surrogate-model-enabled physical-model-free control method. The evolutions of cumulative

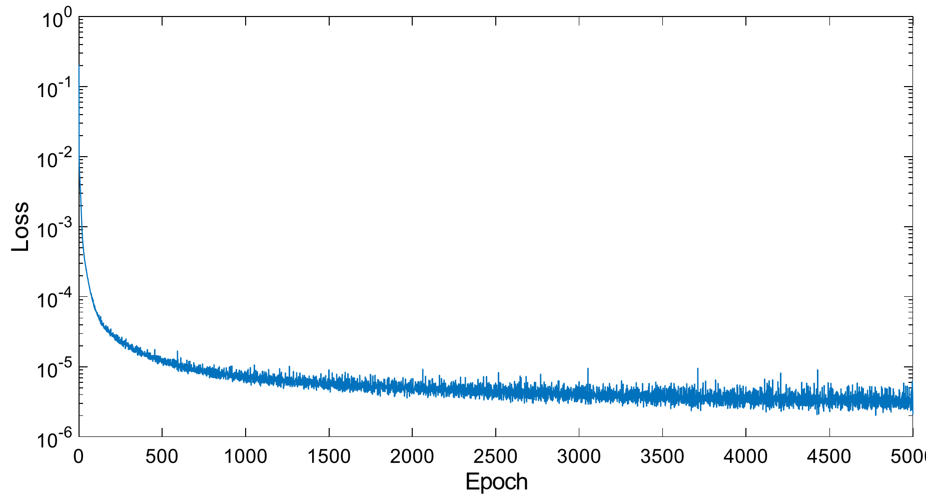

Fig. 3. The loss-epoch curve of the surrogate model during the training process.

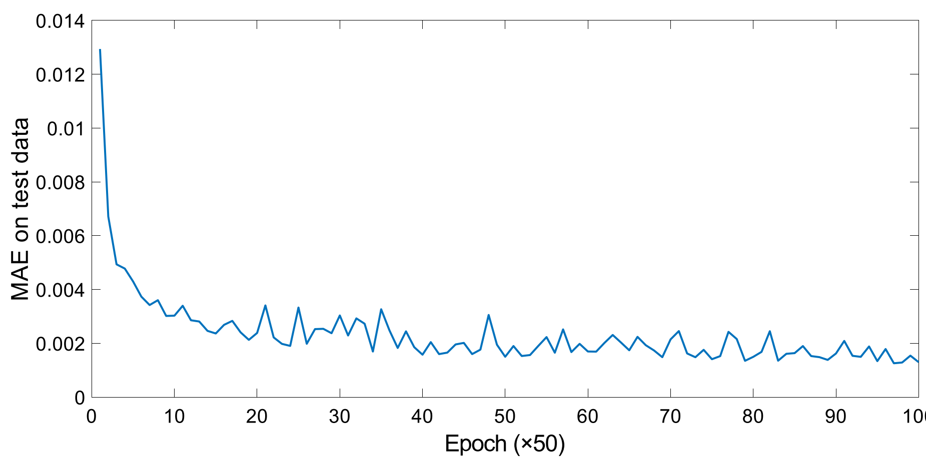

Fig. 4. The evolution of accuracy on test data during the training process.

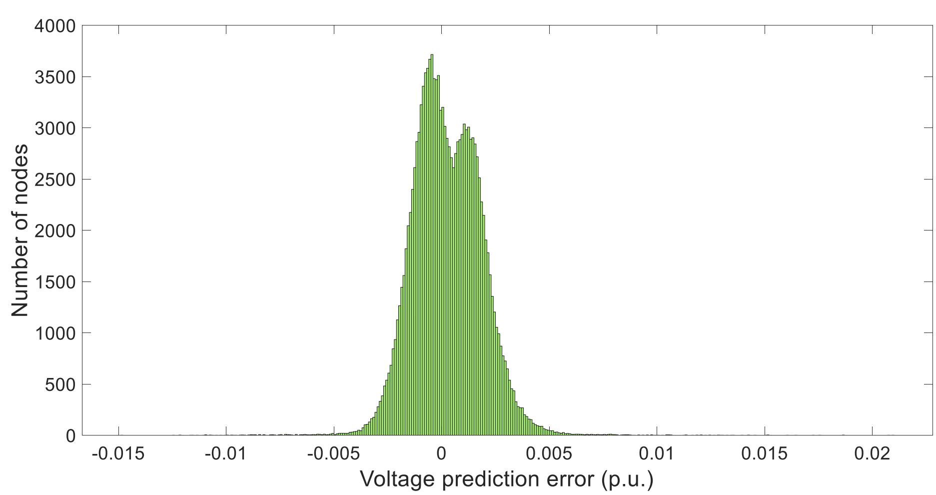

Fig. 5. The distribution of the voltage prediction error for each node.

rewards of various methods on test data during the training procedure are shown in Fig. 7. Evaluations on the test set are carried out every 50 epochs. For better visualization, only the cumulative reward obtained after 15000 episodes are plotted in the figure. The voltage control performance achieved by the proposed method on test data outperforms that of the B-DDPG methods. The results on test data are consistent with those observed during training.

To test the performance of the learned control strategy from the training data, comparative tests are carried out on the test set, which consists of 10 days' data. The voltage profiles using the different

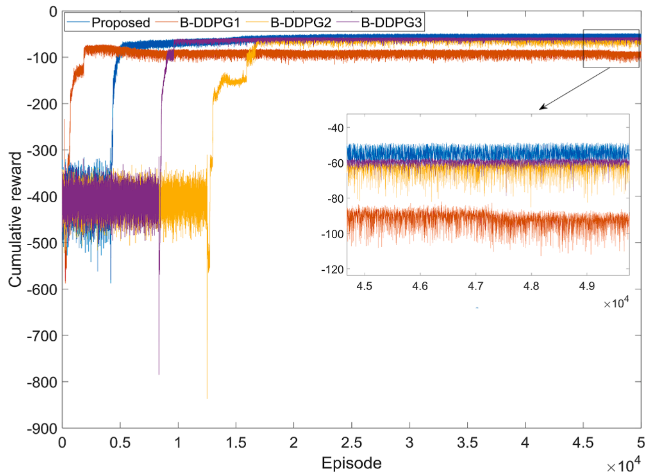

Fig. 6. The changes of the cumulative reward of the proposed and B-DDPG method during the training process.

approaches are shown in Table 5. The methods for comparisons include: 1) **original method** without control; 2) **double deep Q-learning (DDQN) based approach** [44]. Note that for the Q-learning-based algorithm, the actions of various controllable devices must be aggregated to avoid the curse of dimensionality. The action set of the DDQN algorithm contains four variables, which control the reactive power of PVs 1–5, PVs 6–9, the SVCs, and the active power curtailments of all PVs respectively. Each variable is discretized into four values, yielding 256 actions in total. There are two shallow layers of the DNN, the numbers of neurons for which are 400 and 400, respectively; 3) **SP method**, where the PV outputs and load demand are assumed to be subject to a normal distribution. 200 sets of scenarios are generated by Monte Carlo sampling, which is then reduced to 20 representative scenarios; 4) **B-DDPG method**, where the hyper-parameters are the same as the proposed approach except that the memory capacity is set to 200000; 5) **model predictive control (MPC) method**, where rolling optimization is implemented based on the latest forecasting information for the scheduling of SVC and PV inverters. The forecasting step and optimization step is set to 4 and 3, respectively. Only the first step of each scheduling solution is implemented each time; 6) **DDPG method** [38], where **the Z-bus method** [45] **with perfect power flow model** is used to calculate the immediate reward instead of the trained surrogate model during the training of the DRL.

It can be found from Table 5 that when no reactive power control is applied, the maximum voltage rise is beyond the bound. When DDQN, SP, and MPC methods are used, the voltage deviation problem can be suppressed. However, the DDQN method cannot fully utilize the capability of controllable devices because of the aggregation and discretization of actions. Compared with the SP and MPC method, the proposed method and the DDPG can achieve better performance since the control decisions are made based on the latest observation instead of

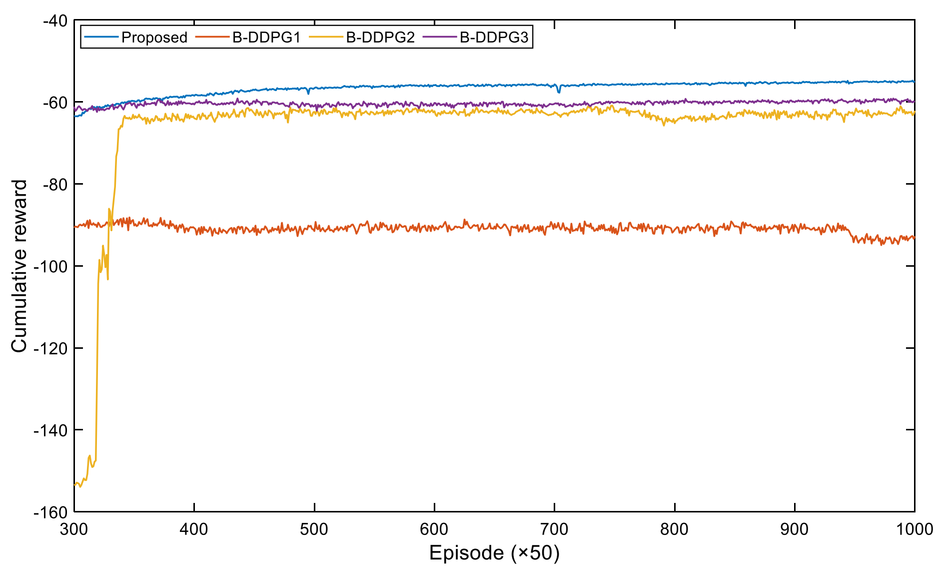

Fig. 7. The evolution of the cumulative reward of the proposed and B-DDPG method on test data during the training process.

Table 5
Voltage deviation of various methods.

Voltage deviation		Original	DDQN	SP	MPC	B-DDPG	Proposed	DDPG
Average deviation		3.64%	1.52%	0.85%	0.86%	0.91%	0.82%	0.81%
Average deviation of each phase	*a*	3.33%	1.29%	0.94%	0.97%	0.97%	0.89%	0.90%
	b	4.15%	1.93%	0.67%	0.68%	0.69%	0.64%	0.63%
	c	3.51%	1.40%	0.91%	0.90%	1.03%	0.90%	0.89%
Max. drop	*a*	4.25%	4.83%	4.29%	4.63%	4.23%	4.26%	4.26%
	b	0.79%	0.85%	3.16%	2.12%	1.83%	1.94%	1.65%
	c	1.84%	4.67%	4.62%	3.76%	3.71%	3.40%	3.49%
Max. rise	*a*	9.74%	4.63%	4.57%	4.57%	4.57%	4.57%	4.57%
	b	8.25%	4.93%	4.57%	4.57%	4.57%	4.57%	4.57%
	c	9.92%	4.57%	4.57%	4.57%	4.57%	4.57%	4.57%
Parameter dependency		–	✓	✓	✓	×	×	✓

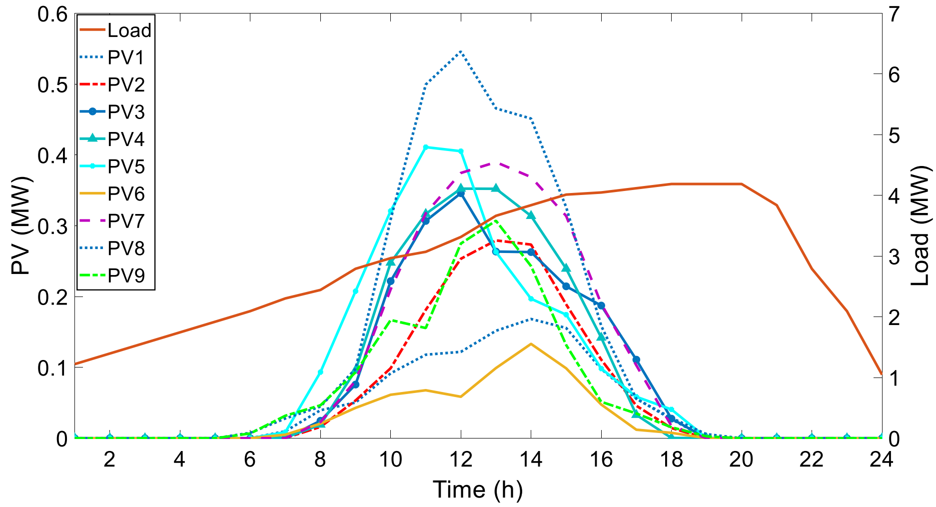

Fig. 8. The PV generations and load demand for a sunny day.

the generated scenarios or forecasting information. It is worth to note that the calculation of the control decisions for SP, MPC, and the training procedures of the DDQN and DDPG depend on the exact knowledge of the parameters and topology of ADN, which are difficult to obtain in practice. By contrast, the proposed approach can obtain performance that is very close to that of the DDPG method without the dependency on the parameters of the ADN by integrating the developed surrogate model. The performance of the proposed method also outperforms that of the B-DDPG approach with an ideal number of experience data, further highlighting the advantages of the proposed method.

A sunny day is selected as a case study to further evaluate the performance of the proposed approach. The PV generations and the load demands of the test day are shown in Fig. 8. The voltage profiles achieved by different methods at $t = 12{:}00$ are shown in Fig. 9. Note that the voltages of the original method, the DDQN, SP, B-DDPG, MPC, the proposed method, and DDPG are obtained by sending the power injections and control decisions to the Z-bus based power flow. The voltages of the surrogate method are calculated by the surrogate model according to the power injection and the control decisions made by the proposed method. It can be observed that the proposed approach and the DDPG method can achieve a better control performance than the DDQN, SP, MPC, and B-DDPG based methods, see buses 43–58 of phases a for example. When the control decisions by the proposed method are implemented, the voltages calculated by the surrogate model are very close to the real value, demonstrating its effectiveness. Thanks to the good forecasting accuracy of the surrogate model, the control strategy learned by interacting with the agent is very similar to that of the DDPG with a perfect power flow model. The PV curtailments of DDQN, SP, B-DDPG, MPC, the proposed method, and DDPG during the test day are 1.96 MW, 0.71 MW, 0.95 MW, 0.65 MW, 0.22 MW, and 0.21 MW, respectively. Since DDQN needs to discretize and aggregate the control decisions of various PVs to avoid the curse of dimensionality, it has much more active power curtailments of PVs than other methods. The proposed method and DDPG curtail less active power of PVs than other methods, demonstrating that they can take full advantage of the reactive power of PVs and SVCs to reduce the voltage deviation.The voltage curve of node 85 of phase a, which suffers from a serious over-voltage problem, is plotted in Fig. 10. The results are consistent with those observed in Table 5 and Fig. 9, demonstrating the effectiveness of the proposed approach.

4.4. Evaluation of impact of modelling errors on the performance of control model

To further illustrate the benefits of the proposed surrogate-model-enabled DRL method, more comparative experiments are conducted in this section. The comparative methods include 1) DDPG-I-1, where an inaccurate physical model of the unbalanced distribution network is available for the training of the DDPG agent. When the training procedure is completed, tests are carried out on the online model (accurate model) to evaluate the performance of the strategy learned by the agent. To simulate the modeling errors, each line parameter is multiplied by a random coefficient ranging from 0.5 to 2 [46]; 2) DDPG-I-2, where an inaccurate model is also utilized during the training of the DRL agent. The modeling errors of this method are achieved by randomly selecting 50% line parameters and multiplying by a random coefficient ranging between 0.5 and 2; 3) DDPG, where the accurate physical model is used for the training of the DDPG agent. Note that DDPG-I-1 and DDPG-I-2 methods simulate the real scenarios since an accurate physical model of unbalance ADN is difficult to get in practice. The performances achieved by various methods on tests data are shown in Table 6. It can be observed from the table that when the modeling errors of all line parameters are considered, the strategy learned by the DDPG-I-1 method during training can adjust the voltages to allowed ranges. However, the average voltage deviations achieved by this method are larger than other methods. This demonstrates that the modeling errors of unbalance ADN harm the performance of the control model. Owing to the smaller modeling errors, the DDPG-I-2 method achieves better control performance than that of the DDPG-I-1 method. The proposed method further enhances the control performance by training a surrogate model instead of using the offline inaccurate model. The DDPG obtains the best voltage control performance but it relies on the accurate network model of the unbalance ADN, which is impossible to obtain in practice. The proposed surrogate-model-enabled DRL method can achieve control performance that is very close to the DDPG method without the requirement of an accurate model. The voltage distributions achieved by different control strategies on the test day are shown in Fig. 11. Note that the voltages of a large number of nodes deviate more than 0.01 p.u. under the control of DDPG-I-1 and DDPG-I-2 methods, illustrating the negative impact of modeling errors. The control performance of the proposed method outperforms that of other methods with an inaccurate model. The results are consistent with that in Table 6.

It can be concluded from the results that the performance of the control model is sensitive to the accuracy of the physical model of unbalance ADN. The proposed approach avoids the negative impact of modelling errors on voltage control performance by training a surrogate model and integrating it with the DRL agent. Comparative tests demonstrate that the voltage control performance achieved by the proposed method is better than other methods with an inaccurate model, illustrating the benefits of the surrogate model.

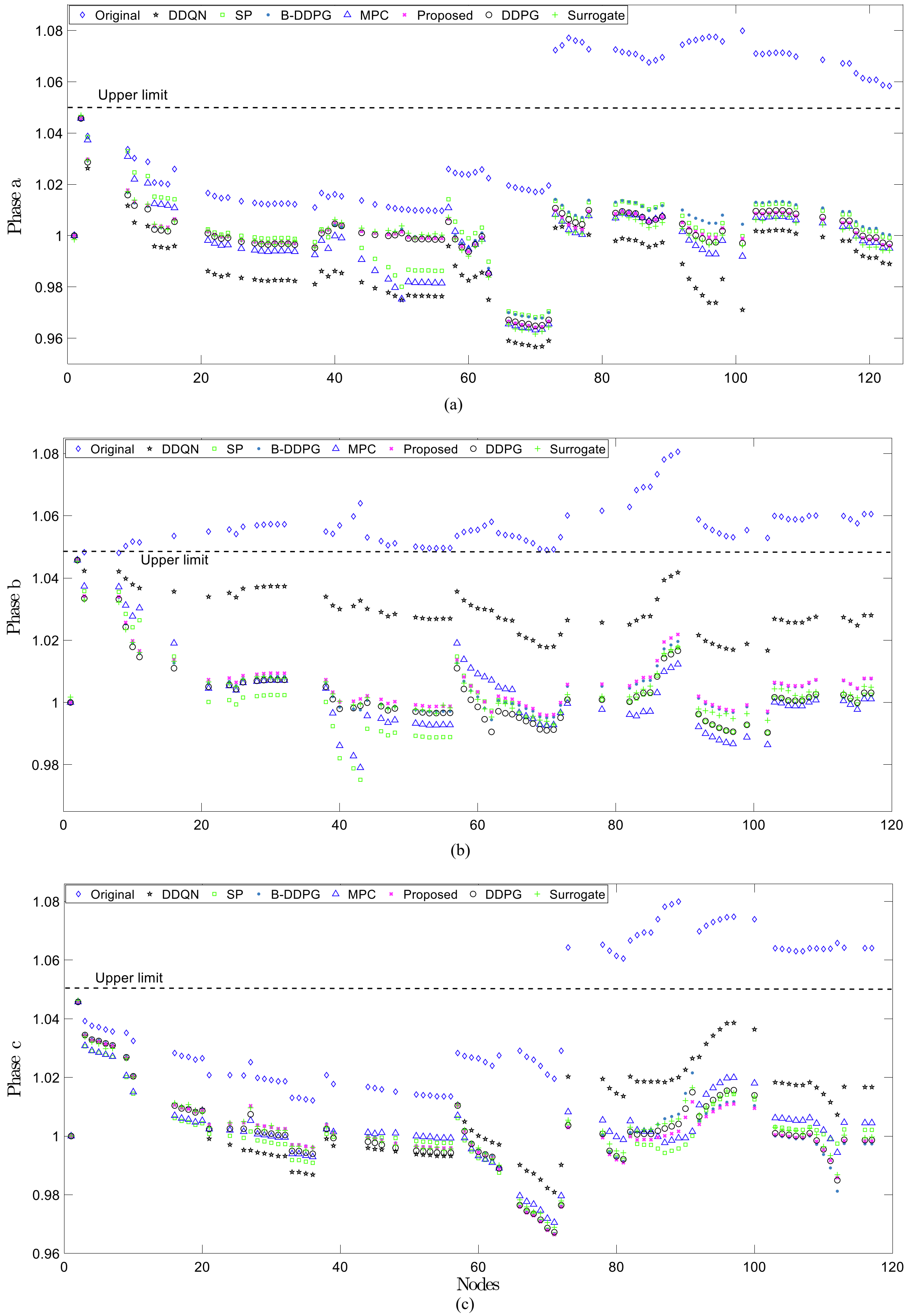

Fig. 9. Voltage profiles of all nodes before and after optimization when t = 12:00. (a) Voltage profiles of phase a. (b) Voltage profiles of phase b. (c) Voltage profiles of phase c.

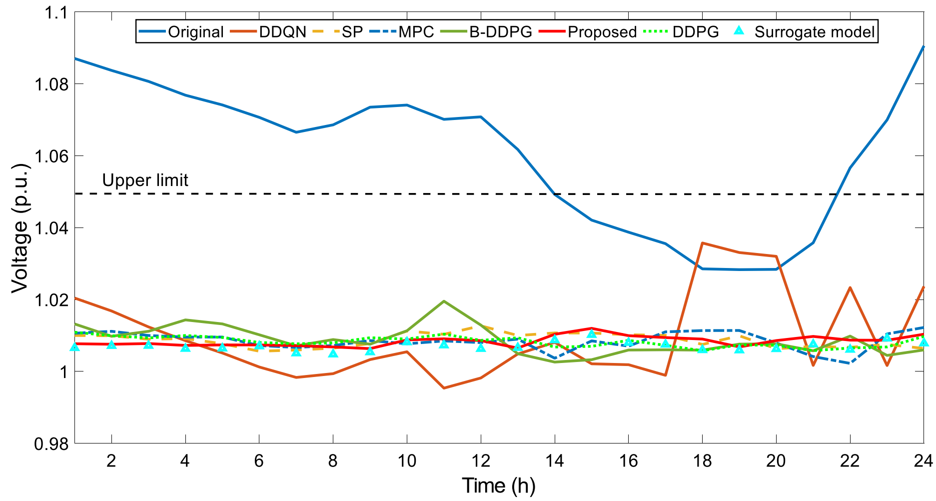

Fig. 10. The voltage profile of node 85 of phase a before and after optimization.

Table 6
Voltage profile achieved by the proposed approach and methods with inaccurate and accurate model.

Voltage deviation		DDPG-I-1	DDPG-I-2	Proposed	DDPG
Average deviation		1.32%	1.09%	0.82%	0.81%
Average deviation of each phase	*a*	1.52%	1.33%	0.89%	0.90%
	b	1.02%	0.76%	0.64%	0.63%
	c	1.39%	1.14%	0.90%	0.89%
Max. drop	*a*	2.76%	3.29%	4.26%	4.26%
	b	1.42%	1.69%	1.94%	1.65%
	c	2.52%	2.86%	3.40%	3.49%
Max. rise	*a*	4.95%	4.57%	4.57%	4.57%
	b	4.57%	4.57%	4.57%	4.57%
	c	4.58%	4.57%	4.57%	4.57%

4.5. Robustness to large stochasticity

More simulations are carried out to demonstrate the advantages of the proposed method in dealing with large PV stochasticity. The calculation times of various methods are first provided in Table 7. The calculation time refers to the time each method takes for calculating each scheduling solution. Since the SP method employs generated scenarios to represent the uncertainties, its calculation burden is relatively high. The MPC method looks for scheduling solutions based on the forecasting information. Therefore, its calculation burden is reduced. By contrast, the DDQN, B-DDPG, proposed, and DDPG methods are reinforcement learning-based methods. They can inform decisions in one millisecond when the training process is finished. This enables them to make real-time decisions based on the latest observation of the ADN.

A rapidly varying PV generation in 1 min owing to the cloud dynamic is tested and the PV output profile is shown in Fig. 12. In this study, the PV output starts to drop from 0.6 MW to 0.3 MW in the 30 s due to the cloud dynamics. Then it starts to rise and takes the 30 s to go back to 0.6 MW. The voltage profile of node 53 of phase *b* before and after optimization is shown in Fig. 13. For the SP and MPC method, a predetermined control decision is used for the voltage control of the whole process. The DDQN, B-DDPG, the proposed, and the DDPG method can provide control decisions in one millisecond. In this case, they provide control decisions every second.

It can be observed that when no control is used for the scheduling of reactive power, there is an over-voltage issue. When the DDQN is applied, the over-voltage is suppressed. However, because the discretization and aggregation of actions hinder the utilization of the reactive power capability of the controllable devices, it suffers from a high voltage deviation. The voltage deviations achieved by the SP and MPC methods are also high owing to the predetermined control decisions cannot provide a flexible reaction to the fast-changing PV output. Since the B-DDPG, the proposed and the DDPG methods can make decisions in milliseconds, they can provide more flexible control decisions based on the latest observations and achieve better voltage regulation performance under large PV output fluctuations. The proposed method achieves better control performance than the B-DDPG method since it learns by interaction with the DNN-based surrogate model instead of using the fixed synthetic experience data stored in replay buffer. It should be emphasized that DDPG is based on a perfect ADN model while our proposed method relies on the surrogate model and DRL algorithm for control. According to the results, we can conclude that although the surrogate model approximations are applied, our proposed method still achieves quite a similar performance as that of DDPG. This means that even without the accurate physical ADN network parameters and topology, our method can be applied in practice. This is the key contribution of this paper and distinguishes it from the existing methods.

4.6. Feasibility of the proposed solutions

Additional tests are carried out to investigate whether the control strategy developed by the DRL agent can always satisfy the voltage constraints. The voltage violation rate during the training process is plotted in Fig. 14. The blue line represents the voltage violation rate of each episode, where 1 indicating that all the decisions the agent made in that episode violate the voltage constraints while 0 represents that all the decisions are feasible solutions. It can be observed from the figure that at the beginning of the training process, the voltage violation rate is very close to 1, demonstrating that the agent has no idea of how to make decisions to satisfy the constraints. The voltage violation rate decreases during training and finally converges to a value that is very close to 0 after 40,000 episodes, indicating that the agent gradually learns the feasible strategy. When the training process is finished, the learned strategy is utilized to inform decisions on the test set. The results in Table 5 show that the maximum drop and arise of voltages obtained by the proposed method are 4.26% and 4.57%, demonstrating that all the decisions are feasible.

Remark: The performance of the control model is sensitive to the accuracy of the physical model of unbalance ADN. However, the estimation of line parameters and topology is difficult since it is affected by many factors. Instead of estimating the parameters of physical model, the proposed method trains a surrogate model which aims to learn the mapping relationship between the power injection and voltage

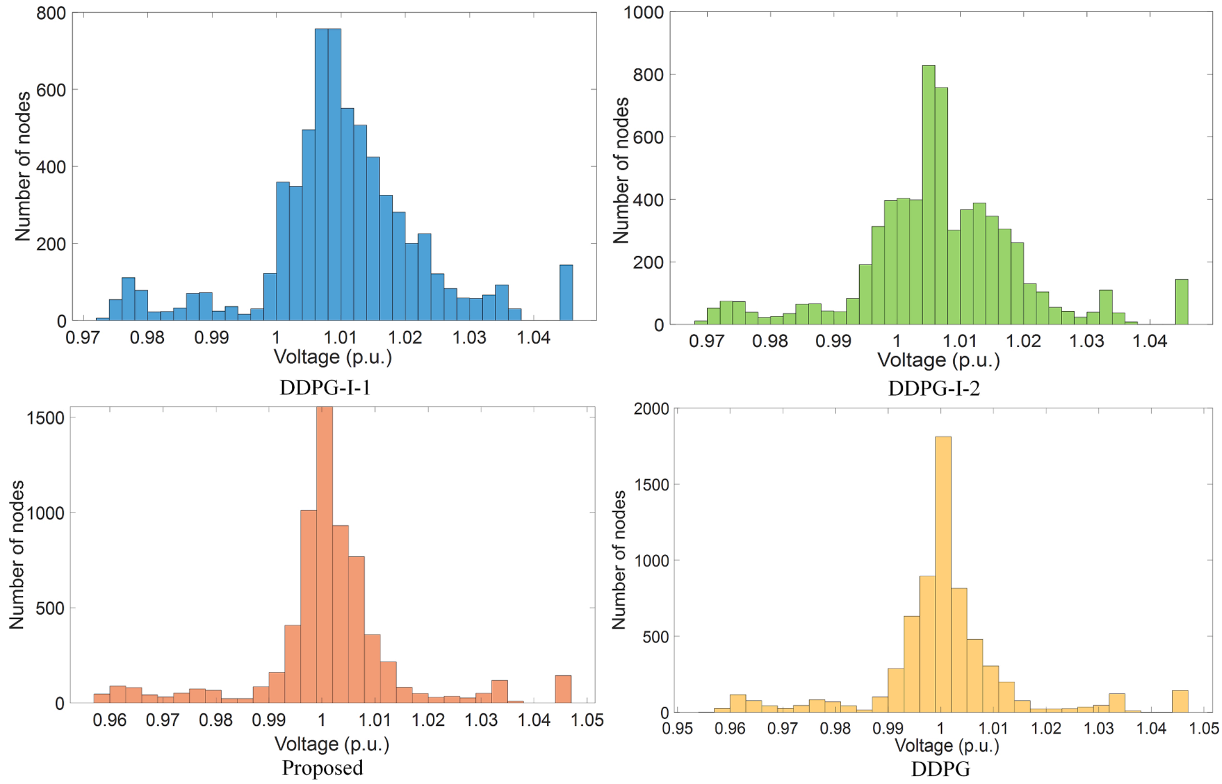

Fig. 11. The voltage distributions obtained by the proposed method, the DDPG method with inaccurate and accurate models, respectively.

Table 7
Calculation time of various methods.

Methods	Original	DDQN	SP	MPC	B-DDPG	Proposed	DDPG
Calculation time (s)	–	0.001	1557	356	0.001	0.001	0.001

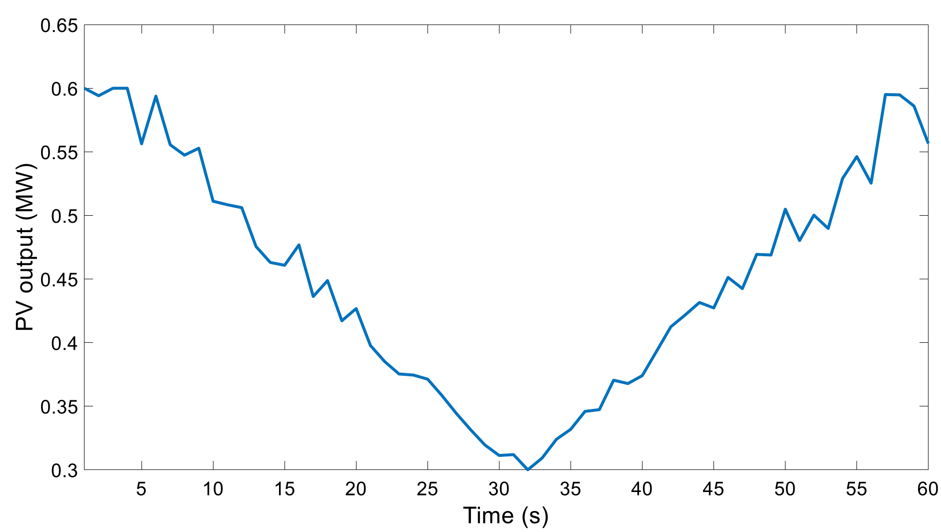

Fig. 12. The PV output profile in the dynamic simulation study.

magnitude of each node, a much easier task. The trained surrogate model is then integrated with the DRL agent and provides reward signal for the development of the voltage control strategy. Numerous tests demonstrate that:

– The proposed DNN-based surrogate model can effectively capture the complex mapping relationship between the power injection and voltage magnitude of each node of distribution network. The high accuracy surrogate model can provide accurate reward signal during the training of the DRL agent. The systematical integration of the DNN-based surrogate model and the DRL agent enables the proposed

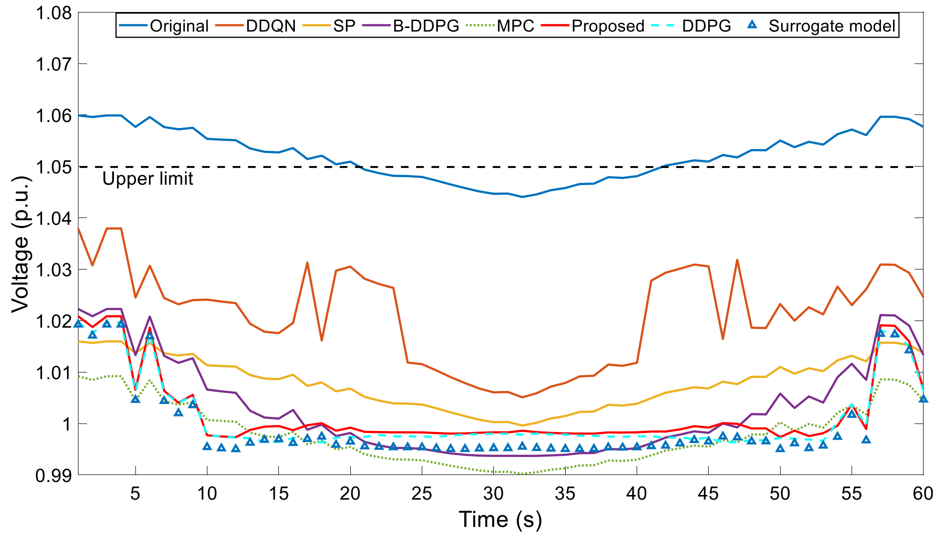

Fig. 13. The voltage profile of node 53 of phase *b* before and after optimization in the dynamic simulation study.

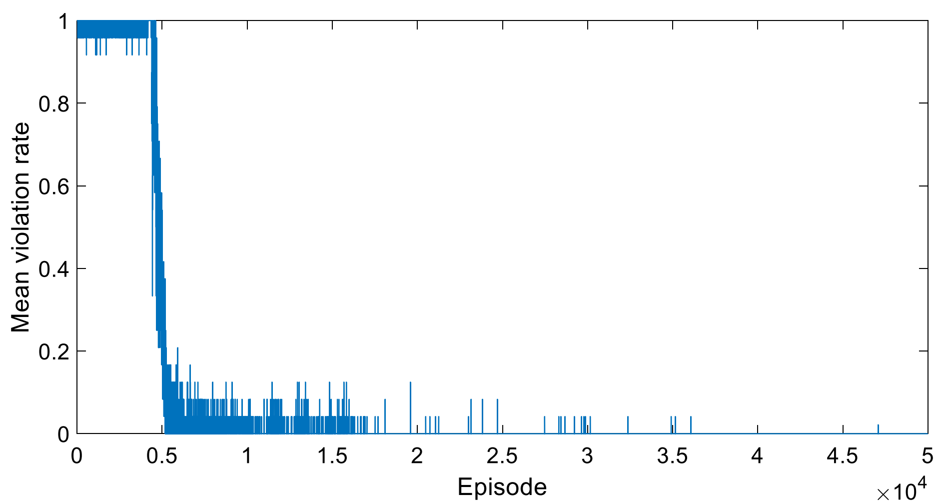

Fig. 14. The voltage violation rate of the proposed method during training.

physical-model-free method to achieve better control performance than methods with inaccurate model and similar control performance as that obtained by DRL method relying on perfect parameters of ADN.

– The proposed method can extract powerful voltage regulation strategy from historical data and inform decisions according to the latest observations in real-time. This allows the proposed method to achieve better control performance than SP method and others. This further enables the proposed method to better deal with fast voltage fluctuations caused by the rapid variation of PV generations.

5. Conclusions

This paper proposes a model-free approach for voltage regulation of three-phase unbalanced distribution network when system parameters and topology are unknown. The proposed approach consists of two components, namely a surrogate model and a deep reinforcement learning control module. The surrogate model is first trained in a supervised manner to learn the complex relationship between the voltage, active and reactive power injections of each node. Then the deep reinforcement learning algorithm is used to learn the voltage regulation strategy from historical data, guided by the immediate reward provided by the surrogate model. The proposed approach can provide voltage control in real-time without the knowledge of system parameters and topology. Numerous comparative tests demonstrate that: 1) the proposed deep neural network-based surrogate model can accurately estimate the voltage magnitude given the active and reactive power injection of each node. The mean absolute error on the test set achieved by the proposed surrogate-model is only 1.3e−3 p.u.; 2) the voltage regulation strategy developed by the proposed deep reinforcement learning agent through interaction with the surrogate model can obtain similar control performance as that achieved by the deep reinforcement learning method with accurate information of the network model. The averaged voltage deviation achieved by the proposed method on the test set is only 1e-4 p.u. higher than the deep reinforcement learning method that relies on an accurate network model; 3) the proposed method can inform real-time decisions according to the latest observations to mitigate fast voltage fluctuations caused by the rapid variation of PV generation, the calculation time is only 0.001 s. The future works include the deployments of an adaptive surrogate model and a meta-learning-based control model, both of which can deal with the topology change of the active distribution network.

CRediT authorship contribution statement

Di Cao: Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft, Supervision. **Junbo Zhao:** Resources, Visualization. **Weihao Hu:** Resources, Visualization. **Fei Ding:** Investigation. **Nanpeng Yu:** Investigation. **Qi Huang:** Resources. **Zhe Chen:** Supervision.

Declaration of Competing Interest

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

Acknowledgements

This research work was supported by National Key Research and Development Program of China (2018YFE0127600).

References

[1] Wang LC, Yan RF, Saha TK. Voltage regulation challenges with unbalanced PV integration in low voltage distribution systems and the corresponding solution. Appl Energy Dec. 2019;256:113927.
[2] Howlader AM, Sadoyama S, Roose LR, Chen Y. Active power control to mitigate voltage and frequency deviations for the smart grid using smart PV inverters. Appl Energy 2020;258:114000.
[3] Ali ES, El-Sehiemy RA, Abou AA, El-Ela KM, Lehtonen M, Darwish MMF. An effective bi-stage method for renewable energy sources integration into unbalanced distribution systems considering uncertainty. Processes 2021;9(3):471.
[4] Hashemi S, Østergaard J, Degner T, Brandl R, Heckmann W. Efficient control of active transformers for increasing the PV hosting capacity of LV grids. IEEE Trans Ind Informat Feb. 2017;13(1):270–7.
[5] Guo Y, Wu Q, Gao H, Huang S, Zhou B, Li C. Double-time-scale coordinated voltage control in active distribution networks based on MPC. IEEE Trans Sustain Energy 2019;11(1):294–303.
[6] Oh SH, Yong TY, Kim SW. Online reconfiguration scheme of self-sufficient distribution network based on a reinforcement learning approach. Appl Energy 2020;280:115900.
[7] Liu M, Procopiou A, Petrou K, Ochoa L, Langstaff T, Harding J, et al. On the fairness of PV curtailment schemes in residential distribution networks. IEEE Trans Smart Grid 2020.
[8] Zeraati M, Golshan MEH, Guerrero JM. Voltage quality improvement in low voltage distribution networks using reactive power capability of single-phase PV inverters. IEEE Trans Smart Grid Sep 2019;10(5):5057–65.
[9] Zeraati M, Golshan MEH, Guerrero JM. Distributed control of battery energy storage systems for voltage regulation in distribution networks with high PV penetration. IEEE Trans Smart Grid Jul. 2018;9(4):3582–93.
[10] Meng K, Dong ZY, Xu Z, Weller SR. Cooperation-driven distributed model predictive control for energy storage systems. IEEE Trans Smart Grid Nov. 2015;6 (6):2583–5.
[11] Emara D, Ezzat M, Abdelaziz AY, Mahmoud K, Lehtonen M, Darwish MMF. Novel control strategy for enhancing microgrid operation connected to photovoltaic generation and energy storage systems. Electronics 2021;10(11):1261.
[12] Stetz T, Marten F, Braun M. Improved low voltage grid-integration of photovoltaic systems in Germany. IEEE Trans Sustain Energy 2012;4(2):534–42.
[13] Xu Y, Dong ZY, Zhang R, Hill DJ. Multi-timescale coordinated voltage/var control of high renewable-penetrated distribution systems. IEEE Trans Power Syst Nov. 2017;32(6):4398–408.
[14] Kekatos V, Wang G, Conejo AJ, Giannakis GB. Stochastic reactive power management in microgrids with renewables. IEEE Trans Power Syst Nov. 2015;30 (6):3386–95.
[15] Zhang C, Xu Y, Dong ZY, Ravishankar J. Three-stage robust inverter-based voltage/ var control for distribution networks with high-level PV. IEEE Trans Smart Grid Jan. 2019;10(1):782–93.
[16] Ding T, Li C, Yang YH, Jiang J, Bie Z, Blaabjerg F. A two-stage robust optimization for centralized-optimal dispatch of photovoltaic inverters in active distribution networks. IEEE Trans Sustain Energy 2017;8(2):744–54.
[17] Li PS, Zhang C, Wu ZJ, Xu Y, Hu M, Dong Z. Distributed adaptive robust voltage/ VAR control with network partition in active distribution networks. IEEE Trans Smart Grid 2020;11(3):2245–56.
[18] Gao HJ, Liu JY, Wang LF. Robust coordinated optimization of active and reactive power in active distribution systems. IEEE Trans Smart Grid Sept. 2018;9(5): 4436–47.
[19] Foggo B, Yu N. Improving supervised phase identification through the theory of information losses. IEEE Trans Smart Grid May 2020;11(3):2337–46.
[20] Wang W, Yu N. Maximum marginal likelihood estimation of phase connections in power distribution systems. IEEE Trans Power Syst Sept. 2020;35(5):3906–17.
[21] Zhao JB, Huang C, Mili L, Zhang Y, Min L. Robust medium-voltage distribution system state estimation using multi-source data. 11th Conference on Innovative Smart Grid Technologies, Washington DC, USA; 2020.
[22] Huang MY, Wei ZN, Zhao JB, Jabr RA, Pau M, Sun G. Robust ensemble Kalman filter for medium voltage distribution system state estimation. IEEE Trans Instrum Meas Jul. 2020;69(7):4114–24.
[23] Wang G, Kekatos V, Conejo AJ, Giannakis GB. Ergodic energy management leveraging resource variability in distribution grids. IEEE Trans Power Syst Nov. 2016;31(6):4765–75.
[24] Karagiannopoulos S, Aristidou P, Hug G. Data-driven local control design for active distribution grids using off-line optimal power flow and machine learning techniques. IEEE Trans Smart Grid Nov. 2019;10(6):6461–71.
[25] Jalali M, Kekatos V, Gatsis N, Deka D. Designing reactive power control rules for smart inverters using support vector machines. IEEE Trans Smart Grid March 2020; 11(2):1759–70.
[26] Arnold DB, Negrete-Pincetic M, Sankur MD, Auslander DM, Callaway DS. Model-free optimal control of var resources in distribution systems: an extremum seeking approach. IEEE Trans Power Syst Sept. 2016;31(5):3583–93.
[27] Yang QL, Wang G, Sadeghi A, Giannakis GB, Sun J. Two-timescale voltage control in distribution grids using deep reinforcement learning. IEEE Trans Smart Grid 2020;11(3):2313–23.
[28] Bui V, Hussain A, Kim H. Double deep Q-Learning-based distributed operation of battery energy storage system considering uncertainties. IEEE Trans Smart Grid 2019;11(1):457–69.
[29] Cao D, Hu W, Zhao J, Huang Q, Chen Z, Blaabjerg F. A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters. IEEE Trans Power Syst Sept. 2020;35(5):4120–3.
[30] Cao D, Zhao J, Hu W, Ding F, Huang Q, Chen Z. Attention enabled multi-agent DRL for decentralized volt-var control of active distribution system using PV inverters and SVCs. IEEE Trans Sustain Energy Jul. 2021;12(3):1582–92.
[31] Cao D, Hu W, Zhao J, Zhang G, Zhang B, Liu Z, et al. Reinforcement learning and its applications in modern power and energy systems: a review. J Mod Power Syst Clean Energy 2020;8(6):1029–42.
[32] Gao Y, Shi J, Wang W, Yu N. Dynamic distribution network reconfiguration using reinforcement learning. In: IEEE SmartGridComm; Oct. 2019. p. 1–7.
[33] Li L, Yang R, Luo D. Efficient fully-offline meta-reinforcement learning via distance metric learning and behavior regularization; [2020, Oct]. [Online] Available: https://arxiv.org/abs/2010.01112.
[34] Gao Y, Wang W, Shi J, Yu N. Batch-constrained reinforcement learning for dynamic distribution network reconfiguration. IEEE Trans Smart Grid Nov. 2020;11(6): 5357–69.
[35] Ding F, Nagarajan A, Chakraborty S, Baggu M. Photovoltaic Impact Assessment of Smart Inverter Volt-VAR Control on Distribution System Conservation Voltage Reduction and Power Quality. Accessed: Jan. 2019. [Online]. Available: https://www.nrel.gov/docs/fy17osti/67296.pdf.
[36] IEEE 1547 Standard for Interconnecting Distributed Resources With Electric Power Systems. Accessed: Jan. 2019. [Online]. Available: http://grouper.ieee.org/groups/scc21/1547/1547_index.html.
[37] Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT press; 2018.
[38] Lillicrap T, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y et al. Continuous control with deep reinforcement learning. Proc 4th Int Conf Learn Represent (ICLR), USA; May. 2016. p. 1–14.
[39] Goodfellow I, Bengio Y, Courville A. Deep learning. The MIT press; 2016.
[40] Tran M-Q, Elsisi M, Mahmoud K, Liu M-K, Lehtonen M, Darwish MMF. Experimental setup for online fault diagnosis of induction machines via promising IoT and machine learning: towards industry 4.0 empowerment. IEEE Access 2021; 9:115429–41.
[41] Lecun Y, Bengio Y, Hinton GE. Deep learning. Nature 2015;521(7553):436.
[42] Cavraro G, Kekatos V. Graph algorithms for topology identification using power grid probing. IEEE Control Syst Lett Oct. 2018;2(4):689–94.
[43] IEEE PES, Distribution Test Feeders; Sep. 2010. [Online]. Available: http://www.ewh.ieee.org/soc/pes/dsacom/testfeeders/index.html.
[44] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: International Conference on Artificial Intelligence; 2016. p. 2094–100.
[45] Bazrafshan M, Gatsis N. Comprehensive modeling of three-phase distribution systems via the bus admittance matrix. IEEE Trans Power Syst Mar. 2018;33(2): 2015–29.
[46] Liu H, Wu W. Two-stage deep reinforcement learning for inverter-based volt-VAR control in active distribution networks. IEEE Trans Smart Grid May 2021;12(3): 2037–47.