Deep Reinforcement Learning Based Approach for Optimal Power Flow of Distribution Networks Embedded with Renewable Energy and Storage Devices

Di Cao, Weihao Hu, *Senior Member*, *IEEE*, Xiao Xu, Qiuwei Wu, Qi Huang, Zhe Chen, *Fellow*, *IEEE*, and Frede Blaabjerg, *Fellow*, *IEEE*

***Abstract*—This study proposes a deep reinforcement learning (DRL) based approach to analyze the optimal power flow (OPF) of distribution networks (DNs) embedded with renewable energy and storage devices. First, the OPF of the DN is formulated as a stochastic nonlinear programming problem. Then, the multi-period nonlinear programming decision problem is formulated as a Markov decision process (MDP), which is composed of multiple single-time-step sub-problems. Subsequently, the state-of-the-art DRL algorithm, i.e., proximal policy optimization (PPO), is used to solve the MDP sequentially considering the impact on the future. Neural networks are used to extract operation knowledge from historical data offline and provide online decisions according to the real-time state of the DN. The proposed approach fully exploits the historical data and reduces the influence of the prediction error on the optimization results. The proposed real-time control strategy can provide more flexible decisions and achieve better performance than the pre-determined ones. Comparative results demonstrate the effectiveness of the proposed approach.**

***Index Terms*—Deep reinforcement learning (DRL), optimal power flow (OPF), wind turbine, distribution network.**

I. Introduction

In the context of the energy shortage, climate change, and environmental protection, the development of clean energy and low-carbon economy, as well as the optimal allocation of energy, is essential [1]. It is an effective way to use sustainable energy by realizing the local consumption of renewable energy in a distribution network (DN). However, renewable energy is affected by natural conditions and has the characteristics of intermittence and uncertainty, presenting challenges to the dispatch and operation of the DN [2].

The optimal power flow (OPF) problems of the DN can be classified into two categories. The first category is deterministic OPF problems. Specific values of the load demand, sustainable generation, and particular network conditions are usually needed to solve this type of problem. Various mathematical approaches [3] and swarm intelligence based approaches are proposed for solving deterministic OPF problems [4], [5]. However, the nonlinear characteristics of these problems (introduced by the constraints of either the network or the devices) make it difficult for the optimization tools to find the global optimum [6]. Evolutionary methods are effective optimization methods when the space of policies is sufficiently small or can be structured and a large amount of time is available for the search [7]. However, power systems have an uncertain nature. It is difficult to implement the intelligence-based methods in the actual operation of power system when considering the uncertainty of the load and the intermittency of renewable energy generation.

The second category is probabilistic OPF (P-OPF) problems. To deal with the uncertainty of the DN, numerous approaches for solving the P-OPF problems have been proposed. References [8], [9] propose stochastic programming based approaches for the optimization of the DN. The stochastic programming based methods assume the knowledge of the distribution of uncertain variables, based on which the scenarios of uncertainty realizations are generated. These methods suffer from a heavy computational burden, as a large number of scenarios must be considered. Moreover, it is difficult to accurately determine the probability distribution of uncertain variables in practice [10]. In contrast to stochastic programming based methods, robust optimization based methods deal with the uncertainty by constructing an uncertainty set and searching the solutions that are robust to all realizations within the set. Robust optimization based methods are proposed for the management of the DN in [11]-[14]. Reference [13] proposes a robust optimization based method that exploits the convex hull tool for the definition of the uncertainty set. Reference [14] proposes a robust quadratic approach for the operation of a smart DN. In the simu-

Manuscript received: July 31, 2020; accepted: April 22, 2021. Date of CrossCheck: April 22, 2021. Date of online publication: June 25, 2021.

D. Cao, W. Hu (corresponding author), X. Xu, and Q. Huang are with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China (e-mail: caodi@std.uestc.edu.cn; whu@uestc.edu.cn; xxiao@std.uestc.edu.cn; hwong@uestc.edu.cn).

Q. Wu is with the Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby 2800, Denmark (e-mail: qw@elektro.dtu.dk).

Z. Chen and F. Blaabjerg are with the Department of Energy Technology, Aalborg University, Aalborg, Denmark (e-mail: zch@et.aau.dk; fbl@et.aau.dk).

DOI: 10.35833/MPCE.2020.000557

lation, the proposed approach achieves better performance than the linearized, nonlinear and quadratically constrained ones. The robust optimization based methods require the obtained solution to be immune to the worst case in the uncertainty set. Thus, the results obtained via these methods are relatively conservative. Chance-constrained methods are also used for the optimization of DN operation [15], [16]. The model predictive control (MPC) algorithm is sometimes used in the two-stage optimization of the management of DNs [17]. However, the performance of the MPC algorithm depends on the accuracy of the prediction of the renewable energy generation and load demand. The past operating experience has not been fully used [18], [19]. The aforementioned methods must resolve a stochastic nonlinear problem partially or completely when a new situation is encountered, which may take some time [20]. Therefore, these methods might not be applicable to real-time control problems. Moreover, these methods greatly depend on accurate information regarding the parameters and topologies of the DN [21]. However, it is hard to obtain the reliable network models in practice.

In recent years, machine learning (ML) has been a popular research topic in computer science. By continuously extracting knowledge from historical data, ML-based methods can generate powerful models to deal with the uncertainty and dynamics of a system without a physical model. The learned models can be generalized to new situations and provide control decisions in real time [22], [23]. Therefore, ML-based methods are promising alternatives with better dynamic performance of real-time optimization of the DN when accurate parameters are unknown [24]. Among the various ML-based methods, reinforcement learning (RL) has the most potential for the optimization of the DN, as it can learn optimal control strategies from historical data without knowing the global optimum [25]. In [26], a *Q*-learning method is used for the energy management of a hybrid electric vehicle. As an effective and famous RL algorithm, *Q*-learning involves learning an action value function, which is a discretized lookup-table matrix. The size of the matrix is determined by the discretized states and actions. When the states and actions are high-dimensional vectors, the sharply increasing matrix size of the action value function makes the convergence difficult. This limits the application of *Q*-learning in practical scenarios with high-dimensional and continuous states and action spaces. To address this problem, [27] proposes the deep RL (DRL) algorithm using a deep neural network (DNN) as the approximator of the action value function. The DNN can take continuous variables as inputs and does not have to discretize the input states. By combining the strong nonlinear approximation ability of the DNN and the decision-making capacity of RL, DRL gives the computer the human-level performance in various complex tasks [28], e. g., play Atari video games and the game Go. In 2016, Google AlphaGo defeated a human champion in chess, which indicates the remarkable potential of DRL.

Various energy management strategies based on the DRL algorithms have been proposed [20], [22]. Reference [20] proposes a deep *Q*-network (DQN) based approach for the management of a battery storage system (BSS) in a microgrid. Simulation results indicate that the proposed approach can deal with the uncertainty of the environment. However, the DQN must discretize the control variables. For optimization problems with a continuous action domain, the discretization of the control variables unusually leads to suboptimal solutions. The deep policy gradient (DPG) based method has been proven effective in the scenarios with high-dimensional and continuous action spaces. Focusing on the building energy optimization problem, [22] proposes a DPG-based method to perform online management of the building energy. The DPG-based method can take multiple actions at the same time and achieve better results than the DQN.

Inspired by recent research, we develop a DPG-based method with continuous action search to solve the P-OPF problem of the DN with renewable energy generation and BSS. The multi-time P-OPF problem is first formulated as a Markov decision process (MDP). Then, the proximal policy optimization (PPO) algorithm, which is the state-of-the-art DPG-based method, is used to solve the MDP, by sequentially considering the influence of the current action on the future. Neural networks (NNs) are used to extract the optimal operation knowledge to cope with the uncertainties from historical data. This model considers the uncertainty of the demand, the initial energy level of the BSS, and the wind power generation. The objective of this model aims to minimize the cost of the power loss by controlling the BSS and the reactive power of the wind turbine under relevant constraints. Comparative experiments are performed using a modified IEEE 33-bus DN to evaluate the performance of the proposed approach. The main contributions of this paper are presented as follows.

First, a real-time energy management strategy for DN based on the DRL algorithm is proposed. The proposed approach embeds operation knowledge extracted from historical data in the DNN to make near-optimal control decisions in real time. The extracted operation knowledge is adaptive to the uncertainty of the system and can be generalized to newly encountered situations. The decision process is similar to recalling the past experience from the memory when a new state is obtained, without resolving the OPF problem. Therefore, the proposed approach can be used for the online optimization of the DN and provide a better response to system dynamics.

Second, the proposed approach decomposes the multi-period decision problem into multiple single-time-step sub-problems, which are sequentially solved while considering their impact on the future. This reduces the computation complexity introduced by the time correlation of the storage devices.

The remainder of this paper is organized as follows. In Section II, the problem formulation is presented. The principle of the proposed approach and the training process are introduced in Section III. The experimental details and the results of a case study are presented in Section IV. Finally, Section V concludes the paper.

II. Problem Formulation

In this section, the mathematical model of the P-OPF problem with wind turbines, load demand, and BSS is presented.

A. Objective Function

The objective of the P-OPF problem is to minimize the cost of power loss. The optimization horizon is 1 day, and the time interval of optimal scheduling is 1 hour. The objective function is formulated as:

$$\min_{P_{bss},Q_{bss},Q_w} F=\min\sum_{t=1}^{T} C_p(t)P_{loss}(t) \tag{1}$$

$$P_{loss}(t)=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}G(i,j)[V_e^2(i,t)+V_f^2(i,t)+V_e^2(j,t)+ V_f^2(j,t)-2(V_e(i,t)V_e(j,t)+V_f(i,t)V_f(j,t))] \tag{2}$$

where F is the total cost of the power loss for an optimization horizon; $P_{loss}(t)$ is the power loss of the DN during hour t; $C_p(t)$ is the electricity price during hour t; $G(i,j)$ is the real component of the complex admittance matrix elements; $V_e(i,t)$ is the real component of the complex voltage at bus i during hour t; $V_f(i,t)$ is the imaginary component of the complex voltage at bus i during hour t; T is the length of one trajectory; and N is the number of nodes in the DN. The control variables are P_{bss}, Q_{bss}, and Q_w, which represent the active power of the BSS, reactive power of the power conditioning system (PCS) of the BSS, and reactive power of the wind turbine, respectively.

B. Constraints

1) Wind Power

The constraints of the active and reactive power of the wind turbine are expressed as [29]:

$$P_w(k,t)=\begin{cases}P_{w,\max}(k) & v_r\le v\le v_{co}\\ P_{w,\max}(k)\left(\dfrac{v}{v_r}\right)^3 & v_{ci}\le v<v_r\\ 0 & \text{otherwise}\end{cases} \tag{3}$$

$$P_w^2(k,t)+Q_w^2(k,t)\le S_w^2(k) \tag{4}$$

where $P_w(k,t)$ is the active power of wind turbine k during hour t; $P_{w,\max}(k)$ is the rated power of wind turbine k; v, v_r, v_{ci}, and v_{co} are the actual speed, rated speed, cut-in speed, and cut-out speed of the wind turbine, respectively; $Q_w(k,t)$ is the reactive power of wind turbine k during hour t; and $S_w(k)$ is the upper bound of the apparent power of wind turbine k. The parameters of the wind turbine are $v_{ci}=4$ m/s, $v_r=14$ m/s, and $v_{co}=24$ m/s.

2) BSS

The BSS consists of a storage unit and a PCS unit. The PCS controls the charging and discharging processes and permits the outputs of active and reactive power, in accordance with the following constraints:

$$P_{bss}^2(k,t)+Q_{bss}^2(k,t)\le S_{PCS,\max}^2(k) \tag{5}$$

$$|P_{bss}(k,t)|\le\bar{P}_{bss}(k) \tag{6}$$

where $P_{bss}(k,t)$ is the active power of BBS k during hour t (when BBS k is charging, $P_{bss}(k,t)$ is a positive value; when it is discharging, $P_{bss}(k,t)$ is a negative value); $Q_{bss}(k,t)$ is the reactive power of BBS k during hour t; $S_{PCS,\max}(k)$ is the upper limit of the apparent power of BBS k; and $\bar{P}_{bss}(k)$ is the charging power limit of BBS k.

The energy balance of the BSS should satisfy (7).

$$\begin{cases}E(k,t+1)-E(k,t)-\eta_{ch}P_{bss}(k,t)=0 & P_{bss}(k,t)>0\\ E(k,t+1)-E(k,t)-\dfrac{P_{bss}(k,t)}{\eta_{dis}}=0 & P_{bss}(k,t)\le 0\end{cases} \tag{7}$$

where $E(k,t)$ is the state of charge (SOC) of BSS k during hour t; and η_{ch} and η_{dis} are the charging and discharging coefficients, respectively. The storage capacity cannot cross the lower or upper bound (20% or 90% of the storage capacity, respectively).

$$E_{\min}\le E(k,t)\le E_{\max} \tag{8}$$

where $E_{\min}$ and $E_{\max}$ are the lower and upper bounds of the SOC of BSS, respectively. Owing to the uncertainty of load demand and renewable energy generation during the intra-day operation, the BSS needs to be flexibly scheduled to cope with the uncertainties in practice. Therefore, the remaining level of BSS is uncertain. In order to get better simulation results of the real circumstance and fully exploit the BSS, the uncertainty of the initial level of BSS is taken into account.

3) Power Flow and Voltage Constraints

The power flow constraints are expressed as:

$$V_e(i,t)\sum_{j=1}^{N}(G(i,j)V_e(j,t)-B(i,j)V_f(j,t))+ V_f(i,t)\sum_{j=1}^{N}(G(i,j)V_f(j,t)+B(i,j)V_e(j,t))+P(i,t)=0 \quad i\in N \tag{9}$$

$$P(i,t)=P_{load}(i,t)-P_w(i,t)+P_{bss}(i,t) \quad i\in N \tag{10}$$

$$V_f(i,t)\sum_{j=1}^{N}(G(i,j)V_e(j,t)-B(i,j)V_f(j,t))- V_e(i,t)\sum_{j=1}^{N}(G(i,j)V_f(j,t)+B(i,j)V_e(j,t))+Q(i,t)=0 \quad i\in N \tag{11}$$

$$Q(i,t)=Q_{load}(i,t)+Q_{bss}(i,t)-Q_w(i,t) \quad i\in N \tag{12}$$

where $B(i,j)$ is the imaginary component of the complex admittance matrix elements; $P(i,t)$ and $Q(i,t)$ are the injection values of the active and reactive power at bus i during hour t, respectively; and $P_{load}(i,t)$ and $Q_{load}(i,t)$ are the active and reactive power of the load demand at bus i during hour t, respectively. Equations (9) and (11) are the active and reactive power flow equations, respectively; and (10) and (12) give the injection values of the active and reactive power, respectively.

The voltage constraint is expressed as:

$$V_{\min}(i)\le V(i,t)\le V_{\max}(i) \quad i\in N \tag{13}$$

where $V(i,t)$ is the voltage at bus i during hour t; and $V_{\min}(i)$ and $V_{\max}(i)$ are the lower and upper bounds of the voltage at bus i, respectively.

The P-OPF problem formulated above is a stochastic nonlinear programming problem with high complexity owing to the network and time domain introduced by the BSS. This study proposes a DRL-based approach to solve this problem, which is described in detail in Section III.

III. Proposed Control Methodology

In this section, the OPF problem is modelled as an MDP first, and then the PPO algorithm is used to solve the MDP. Subsequently, the DNN architecture for function approximation is presented. Finally, the training process of the proposed approach is illustrated in detail.

A. MDP Modelling

The MDP is used to model RL problems. As the optimization of the DN is a sequential decision-making problem, it can be modelled as an MDP with finite time steps. The MDP can be divided into four parts: $\langle S, A, P, R \rangle$.

1) S represents the state set. The state s_t is composed of five parts: $P_{load}(i,t)$, $Q_{load}(i,t)$, $P_w(k,t)$, $E(k,t)$, and $C_p(t)$.

2) A represents the action set. The action a_t is composed of three parts: $P_{bss}(k,t)$, $Q_{bss}(k,t)$, and $Q_w(k,t)$.

3) P represents the probability of a transition to the next state s_{t+1} after action a_t is taken in state s_t. The state transition from s_t to s_{t+1} can be expressed as $s_{t+1}=f(s_t, P_{bss}(k,t), \omega_t)$, where ω_t represents the randomness of the environment. The state transition for the SOC of BSS $E(k,t)$ is controlled by $P_{bss}(k,t)$. This can be denoted explicitly by the equality constraint in (7). Since the wind power generation and load demand for the next hour are not accurately known, the state transitions of $P_{load}(i,t)$ and $P_w(k,t)$ are subject to the environmental randomness. However, it is difficult to accurately model the randomness ω_t in practice. To address this problem, a model-free DRL-based approach is used to learn the transition procedure from historical data, as described in Section III-B.

4) R represents the reward r_t after action a_t is taken in state s_t. A single-step reward r_t is defined as:

$$r_t = P_{loss}(t)C_p(t) + \delta_1 + \delta_2 + p(t) \tag{14}$$

$$p(t) = \begin{cases} -\eta(20\% - E(t)) & E(t) < 20\% \\ 0 & 20\% \le E(t) \le 90\% \\ -\eta(E(t) - 90\%) & E(t) > 90\% \end{cases} \tag{15}$$

where δ_1 is the penalty applied when the voltage exceeds the limit; δ_2 is the penalty applied when the capability limitation of PCS is not satisfied; $p(t)$ is the penalty applied when the upper or lower bound of the storage unit is exceeded; and η is a coefficient. The units of δ_1, δ_2, and η are \$/MWh, thus, the penalty terms have the same measurement term as the cost of the power loss.

At time step t, the agent makes a decision a_t based on the observation of the environment s_t and then obtains a reward r_t. Then, the environment transfers to the next state s_{t+1}. This is an MDP. In the context of the P-OPF, the SOC of BSS is a continuous variable, which is affected by the charging/discharging action performed by the agent. Therefore, when determining a_t, it is reasonable to consider the future reward that the agent obtains after performing action a_t. However, the same reward may not be obtained by the agent the next time, even if the same action is considered, owing to the stochastic nature of the environment (i. e., the uncertainty of wind power generation). Therefore, it is necessary to introduce a discount factor $\gamma \in [0, 1]$ to represent the uncertainty of the environment. The discounted cumulative reward that the agent obtains after action a_t is performed in state s_t is expressed as:

$$R(t) = \sum_{k=0}^{T-t} \gamma^k r_{t+k} \tag{16}$$

The objective of the RL is to learn a policy, which maps the state s_t to the action a_t that can maximize the discounted cumulative reward. By formulating the multi-period optimization problem as an MDP with finite time steps, the problems can be solved sequentially using the DRL algorithm by considering their influence on the future. Instead of solving the multi-period optimization problem by traditional approaches, sequentially solving the MDP helps reduce the computation complexity of the proposed approach. The overall structure of the proposed approach for optimization is illustrated in Fig. 1.

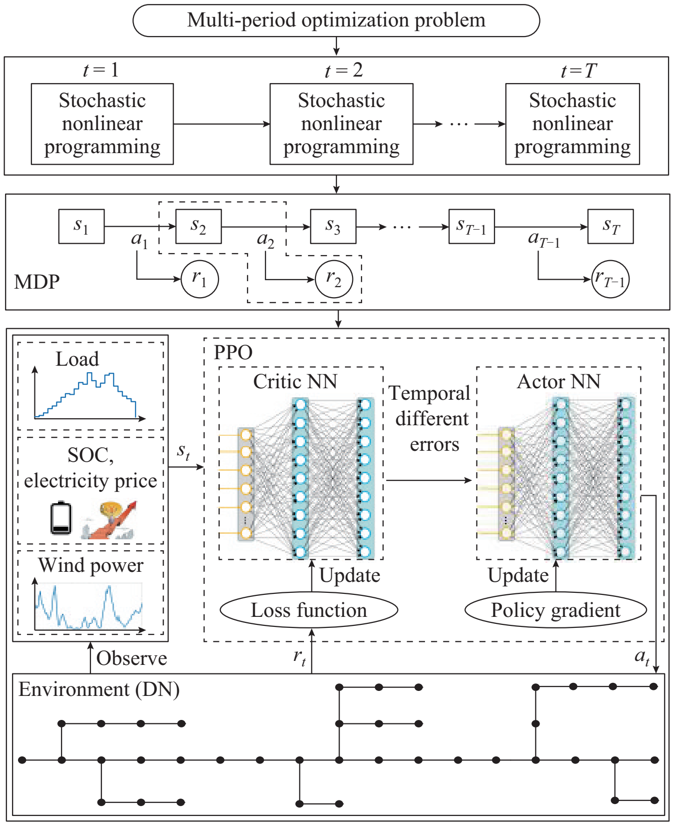


Fig. 1. Overall structure of proposed approach for optimization.

It should be noted that although the introduction of the discount factor reduces the complexity of the proposed approach, the selection of γ requires trial and error process, which is a deficiency of the decomposition.

B. Adopting PPO Algorithm to Solve MDP

PPO is an actor-critic based algorithm (consisting of an actor and a critic). The actor is the policy function that maps the state s_t to the action a_t. The critic is the value function that maps the state s_t to a scalar that measures the quality of the input state.

The actor corresponding to the policy function is parameterized by θ^μ. In traditional policy-based approaches, the parameters are updated by maximizing the reward [7], which is expressed as:

$$\nabla R_{\theta^{\mu}}=\mathbb{E}_{\tau\sim p_{\theta^{\mu}}(\tau)}(R(t)\nabla\lg p_{\theta^{\mu}}(\tau))\approx\frac{1}{N}\sum_{n=1}^{K}\sum_{t=1}^{T}R(t_n)\nabla\lg p_{\theta^{\mu}}(a_{t,n}|s_{t,n}) \tag{17}$$

where $\mathbb{E}(\cdot)$ is the expectation function; K is the number of trajectories; $p_{\theta^{\mu}}(a_{t,n}|s_{t,n})$ is the probability of taking action $a_{t,n}$ in state $s_{t,n}$ under the policy, which is parameterized by θ^{μ}; $\nabla\lg p_{\theta^{\mu}}(a_{t,n}|s_{t,n})$ is the direction that improves the probability of choosing action $a_{t,n}$ in state $s_{t,n}$; and $R(t_n)$ is the reward, which indicates the extent of the probability improvement. Therefore, $\nabla R_{\theta^{\mu}}$ can adjust the strategy in the direction that increases the probability of action with a greater reward value in state s_t.

In (17), since $R(t_n)$ represents the discounted cumulative reward that the agent obtains after state $s_{t,n}$, the parameters of the actor network can only be updated after one episode is completed, which reduces the learning efficiency. To solve this problem, the critic network parameterized by θ^{Q} is introduced. The critic network maps state s_t to a scalar $V^{\pi}(s_t)$, which is the expected cumulative reward that the agent obtains after visiting state s_t under policy π. The $R(t_n)$ in (17) can be replaced with the temporal-difference error, which is given by the value function $A(s_t,a_t)$, as shown in (18):

$$A(s_t,a_t)=r(s_t,a_t)+\gamma V^{\pi}(s_{t+1})-V^{\pi}(s_t) \tag{18}$$

The temporal-difference error indicates the advantage of performing action a_t in state s_t over the expected reward value of all actions. Since $r(s_t,a_t)$ is the immediate reward, the parameter can be updated step by step. The parameters of the value function are optimized by minimizing $L(\theta^{Q})$.

$$L(\theta^{Q})=\mathbb{E}((V^{\pi}(s_t)-y_t)^2) \tag{19}$$

$$y_t=r(s_t,a_t)+\gamma V^{\pi}(s_{t+1}) \tag{20}$$

However, each batch of data can only be used to update the parameter θ^{μ} once, which is a disadvantage of traditional policy gradient methods. To improve the data efficiency and prevent policy updates from becoming too large simultaneously, a clipped objective function is proposed [30]. The parameters of the policy function θ^{μ} are updated by:

$$L^{CLIP}(\theta^{\mu})=\sum_{(s_t,a_t)}\min\left(\frac{p_{\theta^{\mu}}(a_t|s_t)}{p_{\theta^{\mu'}}(a_t|s_t)}A(s_t,a_t),clip\left(\frac{p_{\theta^{\mu}}(a_t|s_t)}{p_{\theta^{\mu'}}(a_t|s_t)},1-\varepsilon,1+\varepsilon\right)A(s_t,a_t)\right) \tag{21}$$

where ε is the clipping rate, which restricts the update range of the new policy in a trusted region; and $\theta^{\mu'}$ is the parameters of the "old" actor, which is in charge of interacting with the environment. The data generated by the "old" actor can be utilized to update the parameters of actor θ^{μ} several times. The clipped function *clip*($\cdot$) helps the PPO algorithm achieve a trade-off among simplicity, sample complexity, and wall-time [30].

C. DNN Architecture for Function Approximation

DNN has a powerful function fitting ability. As reported in [31], an NN can approximate the functions of arbitrary complexity with arbitrary precision. Therefore, NNs are used to fit the value function and policy function in this paper.

In the PPO algorithm, the actor represents the policy function, which maps state s_t to action a_t, and s_t and a_t are the input and output of the policy function, respectively.

$$a_t=z_l^{\mu}(z_{l-1}^{\mu}(\dots z_1^{\mu}(s_t))) \tag{22}$$

$$z_i^{\mu}=g(W_i^{\mu}o_{i-1}^{\mu}+b_i^{\mu})\quad i=2,3,\dots,l \tag{23}$$

where z_i^{μ} is the mapping relationship of the i^{th} layer of the policy function; o_{i-1}^{μ} is the output of the $(i-1)^{\text{th}}$ layer; W_i^{μ} and b_i^{μ} are the weight and bias of the i^{th} layer of the policy function, respectively; and $g(\cdot)$ is the activation function of the neurons.

The critic represents the value function, which maps the state s_t to $V^{\pi}(s_t)$:

$$V^{\pi}(s_t)=z_l^{Q}(z_{l-1}^{Q}(\dots z_1^{Q}(s_t))) \tag{24}$$

$$z_i^{Q}=f(W_i^{Q}o_{i-1}^{Q}+b_i^{Q})\quad i=2,3,\dots,l \tag{25}$$

where z_i^{Q} is the mapping relationship of the i^{th} layer of the value function; o_{i-1}^{Q} is the output of the $(i-1)^{\text{th}}$ layer; W_i^{Q} and b_i^{Q} are the weight and bias of the i^{th} layer of the value function, respectively; and $f(\cdot)$ is the activation function of the neurons.

Therefore, the policy function and value function are parameterized by $\theta^{\mu}=\{W_1^{\mu},b_1^{\mu},W_2^{\mu},b_2^{\mu},\dots,W_l^{\mu},b_l^{\mu}\}$ and $\theta^{Q}=\{W_1^{Q},b_1^{Q},W_2^{Q},b_2^{Q},\dots,W_l^{Q},b_l^{Q}\}$, respectively.

D. Training Process

The training process of the DNN is presented in Algorithm 1. The parameters of the proposed approach can be denoted as $\theta=\{\theta^{\mu},\theta^{\mu'}\theta^{Q}\}$. At the beginning of the training process, the θ of all the NNs are randomly initialized. The parameters of the "old" actor $\theta^{\mu'}$ are copied from θ^{μ}. Then, the algorithm is trained for M episodes to adjust θ. Several actors parameterized by $\theta^{\mu'}$ simultaneously interact with the environment. At the beginning of an episode, each "old" actor obtains a start state s_1 of a day randomly chosen from the training data. At each time step, the actor chooses the action according to the input state s_t. The action is then performed, and the environment transfers to the next state; simultaneously, a reward is obtained. Then, the advantage estimates are calculated using (18). When all the actors finish T time steps, the parameters of the policy network θ^{μ} are updated by:

$$L(\theta^{\mu})=\frac{1}{M}\cdot\sum_{(s_t,a_t)}\min\left(\frac{p_{\theta^{\mu}}(a_t|s_t)}{p_{\theta^{\mu'}}(a_t|s_t)}A(s_t,a_t),clip\left(\frac{p_{\theta^{\mu}}(a_t|s_t)}{p_{\theta^{\mu'}}(a_t|s_t)},1-\varepsilon,1+\varepsilon\right)A(s_t,a_t)\right) \tag{26}$$

$$\theta_{t+1}^{\mu}=\theta_t^{\mu}-\eta_{\mu}\nabla_{\theta^{\mu}}L(\theta^{\mu}) \tag{27}$$

where η_{μ} is the learning rate for the policy network; and M is the mini-batch size. Owing to the introduction of the clipped function, the collected data can be used for updating θ^{μ} several times. Simultaneously, the parameter of the critic network is updated by minimizing the loss $L(\theta^Q)$.

$$L(\theta^Q)=\frac{1}{M}(V^{\pi}(s_t)-y_t)^2 \tag{28}$$

$$\theta^Q_{t+1}=\theta^Q_t-\eta_Q\nabla_{\theta^Q}L(\theta^Q) \tag{29}$$

where η_Q represents the learning rate for the critic network. At the end of each episode, set $\theta^{\mu'}\leftarrow\theta^{\mu}$. When the training is finished, the parameters of the algorithm can be output for real-time optimization of the DN.

```
Algorithm 1: training process of DNN
Input: $\eta_Q$, $\eta_\mu$, $\varepsilon$, $M$, $\gamma$, $T$, $N_a$
Output: $\pi$
Model initialization: randomly initialize critic network $Q(s,a|\theta^Q)$ and
    actor $\mu(s|\theta^\mu)$ with parameters $\theta^Q$ and $\theta^\mu$, and initialize "old" actors
    with parameters $\theta^{\mu'}\leftarrow\theta^\mu$
for $episode=1{:}M$ do
    for $actor=1{:}N_a$ do
        Start state $s_1$ of a random day
        for time step $t=1{:}T$ do
            Select action according to (22), execute $a_t$, obtain re-
              ward $r_t$, and the environment transfers to next state
            Compute advantage estimates $A(s_t,a_t)$ according to (18)
        end for
    end for
    Optimize the parameters of the actor network $\theta^\mu$ according to (26)
      and (27)
    Optimize the parameters of the critic network $\theta^Q$ according to
      (28) and (29)
    Update parameters of "old" actors: $\theta^{\mu'}\leftarrow\theta^\mu$
end for
```

E. Reward Rescaling Based on Clipped Reward Function

Owing to the uncertainty of the environment, the variance of the reward is large. This reduces the accuracy of the value-function estimation and increases the variance of the policy gradient, which may reduce the convergence speed and even lead to a suboptimal policy. To address this problem, a clipped function based reward-rescaling technology is introduced in this paper. The reward sent to the value function is scaled as:

$$r_t=clip\left(\frac{r_t-m}{\sigma},-b,b\right) \tag{30}$$

where m and σ are the mean value and variance of the cumulative discounted reward of an episode, respectively; and $-b$ and b are the lower and upper bounds of reward r_t, respectively. The variance of the rescaled reward is significantly reduced, which helps the value function to learn unbiasedly.

IV. Case Study

In this section, the performance of the proposed approach is analyzed according to numerical results for a DN system. First, the application scenario is presented. Second, the experimental setup is detailed. Third, the training process is described to demonstrate that the algorithm can extract useful operation knowledge from the training data to reduce the cost of power loss. Fourth, a comparison is performed using test data to illustrate the generalization ability of the extracted operation knowledge and the benefits of the proposed approach.

A. Application Scenario

The proposed approach is tested on a modified IEEE 33-bus system to demonstrate the potential for reducing the cost of power loss in the DN. The topology of the DN is shown in Fig. 2. The BSSs are connected to buses 8, 15, 24, and 31. Distributed wind turbines are connected to buses 5, 10, 16, 20, 26, 30, 35, and 36. Bus 1 is selected as the slack bus, and the other buses are PQ buses.

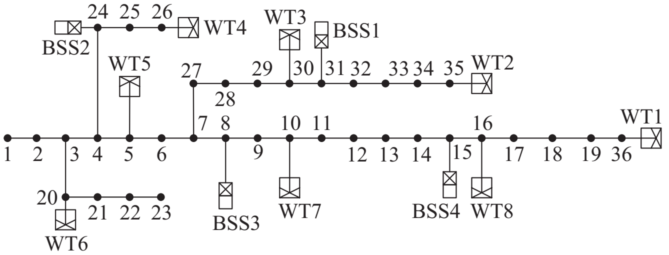


Fig. 2. Topology of DN for case study.

The peak price is 117 \$/MWh and the off-peak price is 65 \$/MWh. The rated power is 500 kW for all the wind turbines. The installed capacity of the BSS is 1000 kWh. The charging and discharging power limit are 300 kW. η_{ch} and η_{dis} are both set as 0.9. The lower and upper bounds of the storage capacity are set as 20% and 90%, respectively. The wind power generation data obtained from western Denmark cover 65 days and are divided into the following two groups. The data of the first 60 days are used as training data (to train the algorithm). The data of the remaining 5 days are used as test data to evaluate the generalization ability of the extracted operation knowledge and the performance of the proposed approach.

B. DNN Architecture and Hyper-parameter Setting

The PPO algorithm is an actor-critic based DRL method that employs an online actor network, a critic network, and a target network. The actor network is a copy of the online actor network. The input of the actor network is the system state s_t, and the output is the action a_t. The input of the critic network is also the system state s_t. The output is the value of the state $V^{\pi}(s_t)$. Both the actor and critic networks have three hidden layers, which have 200, 100, and 100 neurons, respectively. The NNs use the rectified linear unit for all the hidden layers and the output layer of the critic networks. The output layer of the actor network uses both the tanh activation unit and the softplus activation unit. A workstation with an NVIDIA GeForce 1080Ti graphics processing unit and an Intel Xeon E5-2630 v4 central processing unit is used for the training. The DRL algorithm is implemented in Python with TensorFlow, and the power loss is computed in MATLAB. The parameters of the DRL algorithm are presented in Table I.

C. Training Process

The proposed approach and the original PPO algorithm without the clipped reward function are trained off line for 5500 episodes to learn the operation knowledge from the training data.

TABLE I
PARAMETERS OF DRL ALGORITHM

Parameter	Value	Parameter	Value
γ	0.99	ε	0.1
η_μ	10^{-3}	M	32
η_Q	2×10^{-3}	T	24

There are 24 steps in each epoch, which represents one day. The cumulative reward during the training procedure is depicted in Fig. 3. The cumulative reward is shown on a log scale for better visualization. At the beginning of the training, the agent cannot make good decisions and explore the action spaces to achieve more reward information in each state. Through constant interactions with the environment, the proposed approach finally learns a good policy to achieve high cumulative rewards. The proposed approach with the clipped function converges faster than the original PPO algorithm. This is because the clipped function reduces the variance of the reward, thereby reducing the unfavourable influence of the uncertainty of the wind power generation on the approximation of the value function.

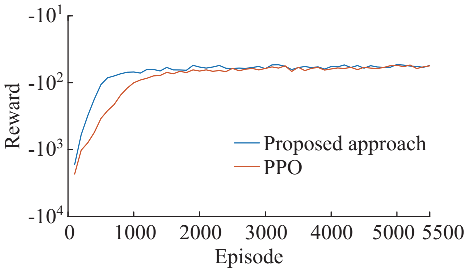


Fig. 3. Cumulative reward during training procedure.

The proportion of satisfied constraints (PSC) and the average cost of the power loss for the training data are shown in Fig. 4. At the beginning of the training, the PSC is almost 0, and the cost of the power loss is high. This is because the agent is unaware of how to make optimal decisions to reduce the cost of the power loss while satisfying the correlated constraints. Therefore, the agent attempts to explore the environment and accumulate experience. The PSC increases sharply until the 1300th episode. At this stage, the cost of the power loss decreases sharply. In this process, the agent learns to make decisions for reducing the cost of the power loss under the correlated constraints.

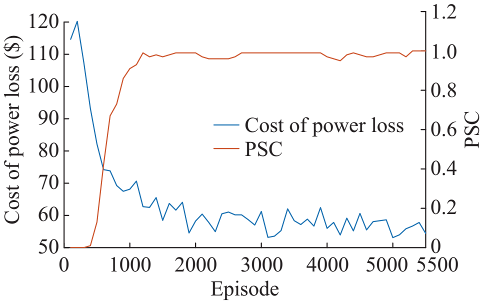


Fig. 4. PSC and cost of power loss during training procedure.

From the 2000th to 5200th episodes, the cost of the power loss is relatively low, while the PSC fluctuates between 0.92 and 1. This indicates that the agent has mastered the skills to reduce the cost of the power loss but sometimes violates the constraints. After approximately the 5200th episode, the PSC is around 1, suggesting that most of the decisions made by the agent satisfy the correlated constraints. This indicates that the proposed approach can extract powerful operating knowledge from training data via the NN to reduce the cost of the power loss under correlated constraints.

D. Comparison Results

1) Experimental Setup

To test whether the knowledge extracted by the NN can be generalized to new situations and to evaluate the performance of the proposed approach, comparative experiments are performed using test data, which cover 5 days. An uncontrolled strategy, the double DQN (DDQN) algorithm, and stochastic programming (SP) are used for comparison. The optimal solution of the proposed approach is the output of the NN, whose parameter is fixed after the training. The DDQN algorithm is an improved version of deep *Q*-learning, which solves the problem of overestimation of the value function when the action dimension is high [32]. The input of the DDQN algorithm is the state s_t, and the output comprises the discrete actions. Owing to the characteristic of the DDQN algorithm, the control variables must be aggregated. There are three types of control variables: P_{bss}, Q_{bss}, and Q_w. Each action is discretized into five values. Therefore, the output layer has 125 neurons, each corresponding to a set of actions. The value functions are approximated by the NNs containing three hidden layers, the neuron numbers of which are 400, 200, and 200, respectively. Note that for the proposed approach, the batteries are uniformly controlled while the wind turbines are controlled separately. For the DDQN and the proposed approach, the uncertainty of the initial SOC of BSS is considered. At the beginning of each episode, the initial SOC of BSS is sampled from Gaussian distribution, the mean and variance of which are 0.5 and 0.1, respectively. The sampled initial SOC of BSS is bounded between 0.2 and 0.9. For the uncertainty modelling of the SP method, it is assumed that the variation of the load demand and the wind power follows a normal distribution. The mean value of the distribution is the forecasted value of the load and wind power. Two hundred sets of scenarios are generated according to the assumed distributions. Then, the number of scenarios is reduced to 20 to reduce the computation burden. Next, the particle swarm optimization algorithm is used to solve the optimization problem.

2) Performance Evaluation

The cost of the power loss with four different methods on five consecutive test days is shown in Fig. 5. As shown, the cost of the power loss varies significantly among the different cases, owing to the variations of the distributed wind energy generation and load demand. However, the proposed approach always has the minimum cost of the power loss. This indicates that the operation knowledge extracted by the NN can be generalized to new situations with various levels of the distributed wind power generation and load demand.

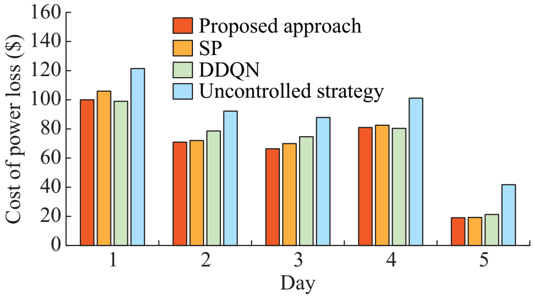


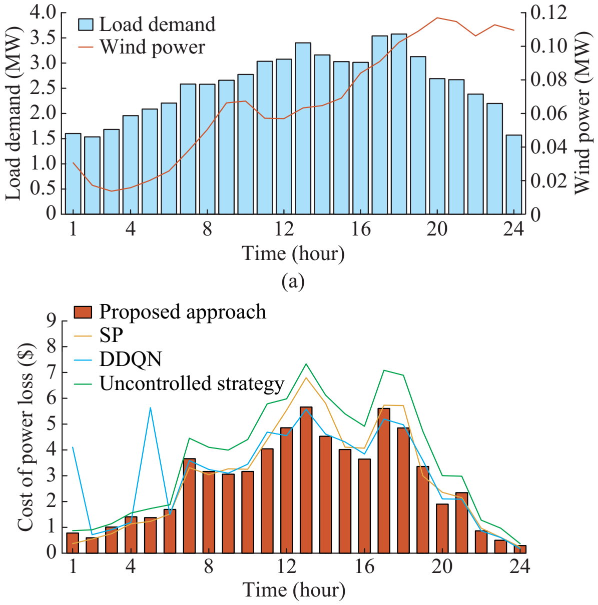
Fig. 5. Cost of power loss with four different methods on five consecutive test days.

The quantitative results are presented in Table II. Compared with the DDQN method, the proposed approach requires no discretization of the actions and avoids information loss during the training procedure; thus, it achieves better results. The proposed approach also achieves better results than the SP method. This may be because the adaptive control strategies learned by the proposed approach during the training procedure are scalable to newly encountered situations. When the training is finished and the algorithm is deployed in a practical system, the proposed approach and the DDQN method can provide the control decisions in a few milliseconds. The decision process is similar to recalling past experience from memory, without resolving the optimization problem. Thus, the proposed approach can provide control decisions based on the latest observed state of the DN. However, the control decisions of SP are pre-determined and cannot be adjusted according to the latest information of the DN. The real-time decisions provided by the adaptive strategies based on the latest information of the DN can yield better results than the pre-determined decisions provided by SP method. This confirms the efficiency of the proposed approach.

TABLE II
QUANTITATIVE RESULTS OF DIFFERENT METHODS

Method	Average cost ($/day)	Improvement (%)
Uncontrolled	88.90	
DDQN	70.78	20.4
SP	69.96	21.3
Proposed	67.48	24.1

The load demand and wind power on a low-wind-speed day and the changes in the cost of the power loss are presented in Fig. 6. Since the optimization horizon is an entire day, no method ensures the global optimum during each hour. It can be observed that the cost of the power loss is high if no control strategy is applied. When the DDQN and SP methods are used, the cost of the power loss is reduced. Compared with the DDQN method, the proposed approach has the continuous action search ability, thus it avoids the information loss and achieves a better control performance. Since the proposed real-time optimization approach makes decisions based on the latest state of the DN, it obtains less cost of the power loss than the pre-determined decisions made by the SP method, i.e., $t=9$-18 hours for example. This is consistent with Fig. 5 and Table II.



(a)

(b)

Fig. 6. Comparison results on low-wind-speed day. (a) Changes in load demand and wind power. (b) Cost of power loss with four different methods.

V. CONCLUSION

The increasing penetration of renewable energy and BSS presents great challenges for the operation of the DN. In this context, we propose a DRL-based approach for the management of the DN under uncertainty. The P-OPF problem is first formulated as an MDP with finite time steps. Then, the PPO algorithm is used to solve the MDP sequentially. NNs are used to obtain the optimal operation knowledge from historical data to deal with the uncertainties. A reward-rescaling function is introduced to reduce the influence of the uncertainty of the environment on the learning process and increase the convergence speed. The operation knowledge extracted from the historical data is scalable to newly encountered situations. When the training is complete, the proposed approach can provide control decisions in real time based on the latest state of the DN, without resolving the OPF problem. Comparative tests confirm that the proposed real-time energy management strategy can provide a more flexible control strategy than the pre-determined decisions provided by the SP method. The proposed DRL-based approach is promising for providing the real-time operation of the DN. Considering that demand response is a promising approach to reduce the power loss by providing consumers with economic incentives, we intend to include it in our future works. The safe DRL-based approach for the optimization of DN while explicitly considering the operation constraints will also be studied in our future works.

REFERENCES

[1] T. Ding, S. Liu, W. Yuan *et al.*, "A two-stage robust reactive power optimization considering uncertain wind power integration in active

distribution networks," *IEEE Transactions on Sustainable Energy*, vol. 7, no. 1, pp. 301-311, Jan. 2016.

[2] A. Gabash and P. Li, "Active-reactive optimal power flow in distribution networks with embedded generation and battery storage," *IEEE Transactions on Power Systems*, vol. 27, no. 4, pp. 2026-2035, Nov. 2012.

[3] M. Aien, M. Rashidinejad, and M. Firuzabad. "Probabilistic optimal power flow in correlated hybrid wind-PV power systems: a review and a new approach," *Renewable & Sustainable Energy Reviews*, vol. 41, pp. 1437-1446, Jan. 2015.

[4] N. Taher, H. Z. Meymand, and H. D. Mojarrad. "An efficient algorithm for multi-objective optimal operation management of distribution network considering fuel cell power plants," *Energy*, vol. 36, pp. 119-132, Jan. 2011.

[5] E. Naderi, H. Narimani, M. Fathi *et al.*, "A novel fuzzy adaptive configuration of particle swarm optimization to solve large-scale optimal reactive power dispatch," *Applied Soft Computing*, vol. 53, pp. 441-456, Apr. 2017.

[6] F. Capitanescu, "Critical review of recent advances and further developments needed in AC optimal power flow," *Electric Power Systems Research*, vol. 136, pp. 57-68, Jul. 2016.

[7] R. S. Sutton and A. G. Barto, *Reinforcement Learning: an Introduction*. Cambridge: MIT Press, 1998.

[8] T. Niknam, M. Zare, and J. Aghaei, "Scenario-based multiobjective volt/var control in distribution networks including renewable energy sources," *IEEE Transactions on Power Systems*, vol. 27, no. 4, pp. 2004-2019, Jul. 2012.

[9] Y. Xu, Z. Dong, R. Zhang *et al.*, "Multi-timescale coordinated voltage/var control of high renewable-penetrated distribution systems," *IEEE Transactions on Power Systems*, vol. 32, no. 6, pp. 4398-4408, Nov. 2017.

[10] D. Bertsimas, E. Litvinov, X. A. Sun *et al.*, "Adaptive robust optimization for the security constrained unit commitment problem," *IEEE Transactions on Power Systems*, vol. 28, no. 1, pp. 52-63, Jan. 2012.

[11] Y. Xu, J. Ma, Z. Dong *et al.*, "Robust transient stability-constrained optimal power flow with uncertain dynamic loads," *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1911-1921, Jul. 2017.

[12] F. Capitanescu and L. Wehenkel, "Computation of worst operation scenarios under uncertainty for static security management," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1697-1705, May 2013.

[13] T. Soares, R. J. Bessa, P. Pinson *et al.*, "Active distribution grid management based on robust AC optimal power flow," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6229-6241, Nov. 2018.

[14] J. F. Franco, L. F. Ochoa, and R. Romero, "AC OPF for smart distribution networks: an efficient and robust quadratic approach," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4613-4623, Sept. 2018.

[15] E. Dall'Anese, K. Baker, and T. Summers, "Chance-constrained AC optimal power flow for distribution systems with renewables," *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 3427-3438, Sept. 2017.

[16] M. Lubin, Y. Dvorkin, and S. Backhaus, "A robust approach to chance constrained optimal power flow with renewable generation," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3840-3849, Sept. 2016.

[17] P. Fortenbacher, A. Ulbig, S. Koch *et al.*, "Grid-constrained optimal predictive power dispatch in large multi-level power systems with renewable energy sources, and storage devices," *IEEE PES Innovative Smart Grid Technologies*, Istanbul, Turkey, Oct. 2014, pp. 1-6.

[18] H. Shuai, J. Fang, X. Ai *et al.*, "Stochastic optimization of economic dispatch for microgrid based on approximate dynamic programming," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2440-2452, May 2019.

[19] H. Shuai, J. Fang, X. Ai *et al.*, "Optimal real-time operation strategy for microgrid: an ADP-based stochastic nonlinear optimization approach," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 2, pp. 931-942, Apr. 2019.

[20] V. Bui, A. Hussain, and H. Kim, "Double deep *Q*-learning-based distributed operation of battery energy storage system considering uncertainties," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 457-469, Jan. 2020.

[21] W. Wang, N. Yu, Y. Gao *et al.*, "Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3008-3018, Jul. 2020.

[22] E. Mocanu, D. Mocanu, P. Nguyen *et al.*, "On-line building energy optimization using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3698-3708, Jul. 2019.

[23] G. Zhang, W. Hu, D. Cao *et al.*, "Deep reinforcement learning-based approach for proportional resonance power system stabilizer to prevent ultra-low-frequency oscillations," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5260-5272, Nov. 2020.

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

[25] D. Cao, W. Hu, J. B. Zhao *et al.*, "A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 4120-4123, Sept. 2020.

[26] X. Qi, G. Wu, K. Boriboonsomsinet *et al.*, "Data-driven reinforcement learning-based real-time energy management system for plug-in hybrid electric vehicles," *Transportation Research Record*, vol. 2572, no. 1, pp. 1-8, Jan. 2016.

[27] V. Mnih, K. Kavukcuoglu, D. Silver *et al.* (2013, Dec.). Playing Atari with deep reinforcement learning. [Online]. Available: https://arxiv.org/abs/1312.5602

[28] V. Mnih, K. Kavukcuoglu, D. Silver *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529-533, Feb. 2015.

[29] G. Kira. "Harvesting the wind: the physics of wind turbines," *Physics and Astronomy Comps Papers*, vol. 2015, pp. 1-41, Apr. 2005.

[30] J. Schulman, F. Wolski, P. Dhariwal *et al.* (2017, Jul.). Proximal policy optimization algorithms. [Online]. Available: https://arxiv. org/abs/1707.06347

[31] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, pp. 251-257, Jan. 1991.

[32] H. Van Hasselt, A. Guez, and D. Silver. (2015, Sept.). Deep reinforcement learning with double *q*-learning. [Online]. Available: https://arxiv.org/abs/1509.06461

Di Cao is currently pursuing the Ph.D. degree in control science and engineering with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include optimization of distribution network and application of machine learning algorithms in power systems.

Weihao Hu received the B.Eng. and M.Sc. degrees in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2007, respectively, and the Ph.D. degree from Aalborg University, Aalborg, Denmark, in 2012. He is currently a Full Professor and the Director of the Institute of Smart Power and Energy Systems, University of Electronics Science and Technology of China, Chengdu, China. His research interests include artificial intelligence in modern power systems and renewable power generation.

Xiao Xu received the B.S. degree in electrical engineering and automation from Qingdao Technological University, Qingdao, China, in 2017. He is currently pursuing the Ph.D. degree in control science and engineering at University of Electronic Science and Technology of China, Chengdu, China. His main research interests include planning and operation optimization of hybrid energy systems.

Qiuwei Wu obtained the Ph.D. degree in power system engineering from Nanyang Technological University, Singapore, Singapore, in 2009. He is currently an Associate Professor at Department of Electrical Engineering, Technical University of Denmark, Copenhagen, Denmark. His research interests include operation and control of power systems with high penetration of renewables, wind power modelling and control, active distribution networks, and operation of integrated energy systems.

Qi Huang received the B.S. degree in electrical engineering from Fuzhou University, Fuzhou, China, in 1996, the M.S. degree from Tsinghua University, Beijing, China, in 1999, and the Ph.D. degree from Arizona State University, Phoenix, USA, in 2003. He is currently a Professor at University of Electronic Science and Technology of China (UESTC), Chengdu, China, the Executive Dean of School of Energy Science and Engineering, UESTC, and the Director of Sichuan State Provincial Lab of Power System Wide-area Measurement and Control. He is a Member of IEEE since 1999. His research interests include power system instrumentation, power system monitoring and control, and power system high performance computing.

Zhe Chen received the B.Eng. and M.Sc. degrees from the Northeast China Institute of Electric Power Engineering, Jilin, China, and the Ph.D. degree from the University of Durham, Durham, UK. He is a Full Professor with the Department of Energy Technology, Aalborg University, Aalborg, Denmark. His research interests are wind energy and modern power systems.

Frede Blaabjerg received the Ph. D. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1995. He was with ABB-Scandia, Randers, Denmark, from 1987 to 1988. He became an Assistant Professor in 1992, an Associate Professor in 1996, and a Full Professor of power electronics and drives in 1998. In 2017, he became a Villum Investigator. He is Honoris Causa at University Politehnica Timisoara, Timisoara, Romania, and Tallinn Technical University, Tallinn, Estonia. His research interests include power electronics and its applications such as in wind turbines, PV systems, reliability, harmonics and adjustable speed drives.