# Sequential Reconfiguration of Unbalanced Distribution Network with Soft Open Points Based on Deep Reinforcement Learning

Ziyang Yin, Shouxiang Wang, and Qianyu Zhao

***Abstract*—With the large-scale distributed generations (DGs) being connected to distribution network (DN), the traditional day-ahead reconfiguration methods based on physical models are challenged to maintain the robustness and avoid voltage off-limits. To address these problems, this paper develops a deep reinforcement learning method for the sequential reconfiguration with soft open points (SOPs) based on real-time data. A state-based decision model is first proposed by constructing a Marko decision process-based reconfiguration and SOP joint optimization model so that the decisions can be achieved in milliseconds. Then, a deep reinforcement learning joint framework including branching double deep *Q* network (BDDQN) and multi-policy soft actor-critic (MPSAC) is proposed, which has significantly improved the learning efficiency of the decision model in multi-dimensional mixed-integer action space. And the influence of DG and load uncertainty on control results has been minimized by using the real-time status of the DN to make control decisions. The numerical simulations on the IEEE 34-bus and 123-bus systems demonstrate that the proposed method can effectively reduce the operation cost and solve the overvoltage problem caused by high ratio of photovoltaic (PV) integration.**

***Index Terms*—Data-driven, distribution network reconfiguration, deep reinforcement learning, distributed generation.**

## NOMENCLATURE

### A. Indices

| | |
|---|---|
| $\phi$ | Index of phases A, B, and C |
| $k, i, j$ | Indexes of node |
| $ij$ | Index of branch from node $i$ to node $j$ |
| $o$ | Index of soft open point (SOP) |
| $s$ | Index of state |
| $t$ | Index of time |

### B. Sets

| | |
|---|---|
| $\Omega_{\mathrm{SOP}}$ | Set of SOPs |
| $A$ | Set of actions |
| $B$ | Set of branches |
| $P$ | Set of transition probabilities |
| $R$ | Set of rewards |
| $S$ | Set of states |
| $T_{\mathrm{SP}}$ | Set of periods |

### C. Parameters

| | |
|---|---|
| $\Delta_t$ | Optimization period |
| $\Delta d_t$ | Number of switching actions during period $t$ |
| $\alpha$ | Temperature coefficient |
| $\beta, \hat{\beta}$ | Parameters of $Q$ target and $Q$ network |
| $\theta, \hat{\theta}$ | Parameters of critic target and critic network |
| $\eta_o$ | Parameters of policy network $o$ |
| $\gamma$ | Attenuation factor |
| $\xi$ | A large number |
| $\lambda^{\mathrm{SP}}$ | Penalty factor (Boolean variable) |
| $\lvert B \rvert$ | Number of samples |
| $c_t^L$ | Electricity price during period $t$ |
| $c_s$ | Per switch action cost |
| $H_l$ | The maximum number of switches in loop $l$ |
| $I_{k,\max}^{\phi}$ | The maximum current of branch $k$ for phase $\phi$ |
| $L$ | Total number of loops |
| $N$ | Total number of buses |
| $N_{\mathrm{SOP}}$ | Total number of SOPs |
| $P_{i,\max}^{\phi,\mathrm{dg}}$, $Q_{i,\max}^{\phi,\mathrm{dg}}$ | The maximum active and reactive output power of distributed generation (DG) connected to bus $i$ for phase $\phi$ |
| $r, X$ | Resistance and reactance |
| $S^{\mathrm{SOP}}$ | Capacity limit of SOP |
| $U_{i,\max}^{\phi}$, $U_{i,\min}^{\phi}$ | The maximum and minimum voltages of bus $i$ for phase $\phi$ |
| $Z$ | Total number of switches |

### D. Variables

| | |
|---|---|
| $\delta_{ij,t}$ | Binary variable representing the opening action of branch $ij$ during period $t$ |

Manuscript received: May 14, 2022; revised: August 8, 2022; accepted: October 8, 2022. Date of CrossCheck: October 8, 2022. Date of online publication: October 28, 2022.

This work was supported in part by the Smart Grid Joint Fund Integration Program of National Natural Science Foundation of China and State Grid Corporation of China (No. U2166202) and National Natural Science Foundation of China (No. 52077149)



Z. Yin, S. Wang, and Q. Zhao (corresponding author) are with Key Laboratory of Smart Grid of Ministry of Education, Tianjin University, Tianjin 300072, China (e-mail: zyyin@tju.edu.cn; sxwang@tju.edu.cn; zhaoqianyu@tju.edu.cn).

DOI: 10.35833/MPCE.2022.000271

| | |
|---|---|
| $\partial_{ij,t}$ | Binary variable representing the closing action of branch *ij* during period *t* |
| $\boldsymbol{A}_t^{\mathrm{SNR}}$ | Sequential network reconfiguration (SNR) action during period *t* |
| $\boldsymbol{A}_t^{\mathrm{SOP}}$ | SOP control strategy during period *t* |
| $A_t^{\mathrm{SP}}$ | Joint control strategy of SNR and SOP during period *t* |
| $A_{t,l}^{\mathrm{SNR}}$ | SNR action on loop *l* during period *t* |
| $I$ | Current |
| $P^{\mathrm{SOP}}$, $Q^{\mathrm{SOP}}$ | Active and reactive power of SOP |
| $P$, $Q$ | Active and reactive power |
| $P^{\mathrm{inj}}$, $Q^{\mathrm{inj}}$ | Active and reactive injection power |
| $P^{\mathrm{sub}}$, $Q^{\mathrm{sub}}$ | Active and reactive output power of generator connected to the node |
| $P^{\mathrm{dg}}$, $Q^{\mathrm{dg}}$ | Active and reactive output power of DG connected to node |
| $P^{\mathrm{load}}$, $Q^{\mathrm{load}}$ | Active and reactive demand power |
| $\boldsymbol{P}^{\mathrm{inj}}$, $\boldsymbol{Q}^{\mathrm{inj}}$ | Three-phase active and reactive injection power |
| $PF_{\mathrm{dg}}$ | Power factor of DG |
| $S^{\mathrm{SP}}$ | Bus injection power |
| $\boldsymbol{S}_{t+1}^{\mathrm{SP}}$ | Bus injection power during period $t+1$ |
| $U$ | Voltage |
| $z_{ij,t}$ | Binary variable denoting state |

## I. Introduction

DISTRIBUTION network reconfiguration (DNR) is an effective way to optimize distribution network (DN) operation. DNR optimizes the operation state of the DN by controlling the sectional switch or tie switch and ensures that the optimization results satisfy the operational constraints [1]. DNR can be divided into static DNR [2]-[5] and dynamic DNR. The static DNR is a single-stage decision method which is mainly used to optimize the state of the switch without changing the load condition of the DN. The latter is based on the dynamic optimization of the DN according to the load change and the operation constraints of the DN during different phases. In addition, sequential decision-making (SDM) is a time-sequential multi-stage optimization problem, where a controller can interact with system to obtain various sequential decisions (strategies) to maximize gains or minimize losses [6]. The SDM problem is more complex than a series of multiple independent decision problems since the controller considers the long-term effects of its decisions [7]. Thus, according to the definition of SDM, dynamic DNR can also be called sequential network reconfiguration (SNR) or multi-stage reconfiguration, which is more suitable than static DNR for the requirements of the actual operation of the DN [8].

Considering the factors such as switching cost and surge current of closing loop, it is impossible for the tie switch to be disconnected frequently. Therefore, the traditional DNR is difficult to realize real-time topology adjustment. However, the soft open points (SOPs) can change the transmission power in real time, adjust the operating status, and realize the flexible interconnection between feeders [9]. Compared with traditional switch operation, SOP can control the power flow accurately and flexibly. Nevertheless, considering the high investment of SOP, it cannot completely replace the tie switch in the short term [10]. Therefore, it is worthwhile to investigate SNR considering SOP (SP).

The SP is a typical mixed-integer nonlinear programming, and the main solution methods include meta-heuristic algorithm (MHA) and mixed-integer programming (MIP). MHA is the product of the combination of random algorithm and local search algorithm, i.e., particle swarm optimization [10], [11], simulated annealing algorithm [12], and grey wolf optimizer [13]. However, the computational burden of MHA is usually too heavy and cannot be used for real-time decision-making. MIP is a more popular method, which uses mathematical models to describe the problem of DNR, and then obtains the optimization results through some mathematical optimization methods, like second-order cone programming [14]-[16] and mixed-integer linear programming [17]. However, MIP still encounters significant challenges when addressing mixed-integer optimization with large numbers of integer variables [18]. Furthermore, existing control algorithms are typically determined offline, which are less optimized and unable to adapt to unknown system changes [19].

Recently, with the development of artificial intelligence technology, power system dispatching methods based on historical data and deep reinforcement learning (DRL) have attracted researchers' attention [20], [21]. DRL formulates the optimization problem considering the uncertainty as a random dynamic program with unknown state transition probability [22]. In [22], a DRL-based voltage control method is proposed, making voltage control strategies according to the real-time system conditions. In practice, the system operators expect to make decisions based on the real-time state with consideration of uncertainty in the future [23]. Therefore, in [23], a Marko decision process (MDP) based DNR method is proposed, using a dynamic programming approach to realize real-time decision-making. In [1], a batch-constrained DRL algorithm without the interaction with DN for the dynamic DNR problem is proposed, using the historical reconfiguration data to learn the SNR strategy. However, the DNR optimization alone is difficult to alleviate the system overvoltage problem caused by distributed generation (DG) with high penetration. And DNR decision is a discrete variable and SOP control result is a continuous variable. The existing DRL-based DNR method cannot solve the joint optimization problem of SOP and DNR. Therefore, it is necessary to study the DRL-based SP problem.

To address the challenges mentioned earlier, a DRL method for interaction with DN is proposed to solve the SP problem, which formulates the SP as a decision-making problem with multi-dimensional action space to minimize the operation cost. The SP model is first converted to an SP based on the Marko decision process (SP-MDP) model to construct a real-time decision model. The bus injection power is used as the state quantity, and the SP optimization strategy is taken as the action quantity. Then, a DRL framework including branching double deep $Q$ network (BDDQN) and multi-poli-

cy soft actor-critic (MPSAC) algorithm is proposed to learn the SP control strategy with the SP-MDP model. Furthermore, the proposed method is evaluated on the IEEE 34-bus system and IEEE 123-bus system with high photovoltaic (PV) penetration. Numerical study results show that the proposed DRL method can successfully learn the SP control strategy and reduce system operation cost.

The significant contributions of this paper are listed below.

1) A DRL-based SNR and SOP joint optimization method is proposed, which constructs the state-based SP decision model with MDP theory, obtains decision results in milliseconds, and improves system operation economics compared with DNR.

2) A DRL framework is proposed including BDDQN for learning reconfiguration strategy by multi-dimensional action-value function and MPSAC for learning SOP control strategy through multi-policy network collaboration, which has better learning stability and performance than traditional DRL algorithm.

3) The proposed method uses the pre-trained BDDQN-MPSAC (BD-AC) agent and real-time bus injection power collected by the SCADA system or phase measurement unit (PMU) system to make optimization decisions. Thus, the influence of DG and load uncertainty on SP decision-making has been reduced to the most extent.

The remainder of this paper is organized as follows. The reinforcement learning modeling for SNR is presented in Section II. A DRL-based SP-MDP solution model is formulated in Section III. The case study is presented in Section IV and the conclusions are shown in Section V.

## II. Reinforcement Learning Modeling for SNR

In this section, the traditional mathematical model of the unbalanced SP is firstly introduced. Then, according to the characteristics of the DN, the SNR problem is converted into the SNR-MDP.

### A. Function and Mathematical Model of SOP

According to different control methods, SOP can be divided into three types: unified power flow controller, static synchronous series compensator, and back-to-back voltage source converter (B2B-VSC). This paper takes B2B-VSC as an example to explain the function and control mode of SOP in the DN, as shown in Fig. 1.

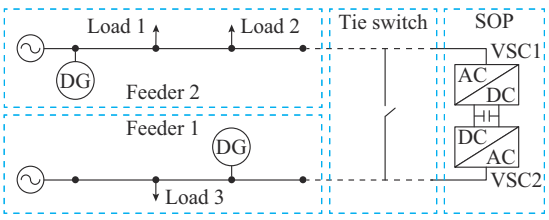


Fig. 1. Function and control mode of SOP in DN.

B2B-VSC can precisely regulate the active power transmitted between two feeders and provide reactive power support. The variables for SOPs consist of the three-phase active outputs $P_{\phi}$ of the converter VSC1 and three-phase reactive power outputs of two converters. Assume that the active power of the two converters is equal, i.e., the active power output of VSC2 is $-P_{\phi}$. The three-phase reactive power output of two converters is not affected by each other due to the DC isolation, so it only needs to satisfy the capacity constraints of each converter [24]. If the three-phase SOP consists of three single-phase SOP modules, the three-phase power of the SOP can be controlled independently. We take the $PQ$-$V_{dc}Q$ control of B2B-VSC to illustrate the mathematical model of SOP [25] with the following constraints.

*1) Active Power Constraint for SOP*

$$P_{o,i,t}^{\mathrm{SOP},\phi}+P_{o,j,t}^{\mathrm{SOP},\phi}=0 \quad o\in\Omega_{\mathrm{SOP}} \tag{1}$$

*2) Capacity Constraints for SOP*

$$\begin{cases}\sqrt{(P_{o,j,t}^{\mathrm{SOP},\phi})^2+(Q_{o,j,t}^{\mathrm{SOP},\phi})^2}\le S_{o,j}^{\mathrm{SOP},\phi}\\ \sqrt{(P_{o,i,t}^{\mathrm{SOP},\phi})^2+(Q_{o,i,t}^{\mathrm{SOP},\phi})^2}\le S_{o,i}^{\mathrm{SOP},\phi}\end{cases} \tag{2}$$

The operation efficiency of SOPs can reach 98% [8], [26]. Thus, the power losses of SOPs are ignored in this paper.

### B. Mathematical Model of SP

The objective function of SP is to minimize the operation cost of DN, including the energy loss and switch action cost. Note that $\Delta_t$ is the optimization period, which is equal to 1 hour.

$$f=\min\sum_{\phi}\sum_{t=1}^{|T_{\mathrm{SP}}|}c_t^L\left(\sum_{ij\in B}(I_{ij,t}^{\phi})^2r_{ij}^{\phi}\right)\Delta_t+\sum_{t=1}^{|T_{\mathrm{SP}}|}c_s\Delta d_t \tag{3}$$

The decision variables are network topology and the three-phase control strategy of SOP. While optimizing the objective function, the following constraints need to be met.

*1) Power Balance Constraints*

The distribution load flow equations [27] are used to ensure the power balance:

$$\begin{cases}\sum_{ki\in B}(P_{ki,t}^{\phi}-(I_{ki,t}^{\phi})^2r_{ki}^{\phi})-\sum_{ij\in B}P_{ij,t}^{\phi}=P_{i,t}^{\phi,\mathrm{inj}}\\ \sum_{ki\in B}(Q_{ki,t}^{\phi}-(I_{ki,t}^{\phi})^2X_{ki}^{\phi})-\sum_{ij\in B}Q_{ij,t}^{\phi}=Q_{i,t}^{\phi,\mathrm{inj}}\end{cases} \tag{4}$$

$$(U_{j,t}^{\phi})^2=(U_{i,t}^{\phi})^2-2(P_{ij,t}^{\phi}r_{ij}^{\phi}+Q_{ij,t}^{\phi}X_{ij}^{\phi})+(I_{ij,t}^{\phi})^2((r_{ij}^{\phi})^2+(X_{ij}^{\phi})^2) \tag{5}$$

$$(I_{ij,t}^{\phi})^2(U_{ij,t}^{\phi})^2=(P_{ij,t}^{\phi})^2+(Q_{ij,t}^{\phi})^2 \quad \forall ij\in B \tag{6}$$

$$\begin{cases}P_{i,t}^{\phi,\mathrm{inj}}=P_{i,t}^{\phi,\mathrm{sub}}+P_{i,t}^{\phi,\mathrm{dg}}-P_{i,t}^{\phi,\mathrm{load}}\\ Q_{i,t}^{\phi,\mathrm{inj}}=Q_{i,t}^{\phi,\mathrm{sub}}+Q_{i,t}^{\phi,\mathrm{dg}}-Q_{i,t}^{\phi,\mathrm{load}}\end{cases} \tag{7}$$

*2) Bus Voltage Constraints*

$$U_{i,\min}^{\phi}\le U_{i,t}^{\phi}\le U_{i,\max}^{\phi} \tag{8}$$

To ensure the power quality of DN, the bus voltage needs to be limited within a safe range.

*3) Branch Power and Current Constraints*

The branch power and current need to be limited within a safe range during DNR.

$$\begin{cases}|S_{ij,t}^{\phi}|\le z_{ij}S_{ij,\max}^{\phi}\\ |I_{ij,t}^{\phi}|\le z_{ij}I_{ij,\max}^{\phi}\end{cases} \quad ij\in B \tag{9}$$

*4) Switch Action Constraints*

The frequent action of a switch will shorten its life span. Therefore, limiting the number of switch actions is necessary to minimize the switching loss while reducing the network loss.

$$\begin{cases} z_{ij,t-1}-z_{ij,t}\le\delta_{ij,t} \\ z_{ij,t}-z_{ij,t-1}\le\partial_{ij,t} \\ \sum\limits_{t\in T}\sum\limits_{ij\in B}(\alpha_{ij,t}+\partial_{ij,t})\le 2S_{\max} \end{cases} \tag{10}$$

*5) Topological Constraint*

The DN must have radial topology with all the buses energized.

$$\sum_{(i,j)\in B} z_{ij,t}=N-1 \tag{11}$$

*6) DG Constraints*

The DG operation constraints can be expressed as:

$$\begin{cases} Q_{i,t}^{\phi,\mathrm{dg}}=P_{i,t}^{\phi,\mathrm{dg}}\tan(\arccos(PF_{\mathrm{dg}})) \\ \sqrt{(P_{i,t}^{\phi,\mathrm{dg}})^2+(Q_{i,t}^{\phi,\mathrm{dg}})^2}\le S_{\phi,i}^{\mathrm{DG}} \end{cases} \tag{12}$$

Then, combining the objective function (1) and constraints (2)-(12), we propose a state-based SP decision optimization model based on MDP theory.

*C. Application of MDP in SP*

RL aims to learn the optimal policy through the interaction process between the agent and the environment. An RL problem can be modeled with MDP, which is a standard formalism for solving SDM problems based on Markov process theory [28]. It can be expressed as:

$$MDP\Leftrightarrow\langle S,A,P,R,\gamma\rangle \tag{13}$$

The detailed introduction of MDP to solve SDM problems can be found in [29]. According to the above objective function and constraints in the SP model, the radical factors which can affect the optimal SP solution are the bus injection power in each period. Therefore, the set of bus injection power in each period is defined as a state set $S^{\mathrm{SP}}$, which can be expressed as:

$$\boldsymbol{S}^{\mathrm{SP}}=\{\boldsymbol{P}_t^{\mathrm{inj}},\boldsymbol{Q}_t^{\mathrm{inj}}\}\quad \boldsymbol{P}_t^{\mathrm{inj}},\boldsymbol{Q}_t^{\mathrm{inj}}\in\mathbb{R}_{1\times 3N},t\in T_{\mathrm{SP}} \tag{14}$$

Afterward, the mixed-integer action set can be defined as a combination of SNR and SOP control strategies.

$$\boldsymbol{A}^{\mathrm{SP}}=\{\boldsymbol{A}_t^{\mathrm{SNR}},\boldsymbol{A}_t^{\mathrm{SOP}}\}\quad \boldsymbol{A}_t^{\mathrm{SNR}}\in\mathbb{R}_{1\times L},\boldsymbol{A}_t^{\mathrm{SOP}}\in\mathbb{R}_{1\times 12N_{\mathrm{SOP}}},t\in T_{\mathrm{SP}} \tag{15}$$

To accommodate the radial operating characteristics of the DN, the $\boldsymbol{A}_t^{\mathrm{SNR}}$ is coded according to the position of action switch in the fundamental loop. Take the modified IEEE 34-bus system in Fig. 2 as an example [18]. The fundamental loop of the IEEE 34-bus system is written as:

$$\boldsymbol{H}=\begin{cases}[16\quad 17\quad 20\quad 25\quad 27\quad 29\quad 30\quad 37]\quad \text{loop 1}\\ [13\quad 14\quad 15\quad 25\quad 27\quad 36]\quad \text{loop 2}\\ [3\quad 5\quad 6\quad 7\quad 34]\quad \text{loop 3}\end{cases}$$

If the action switches during period $t$ are s5, s25, and s27, $\boldsymbol{A}_t^{\mathrm{SNR}}=[2,4,5]$.

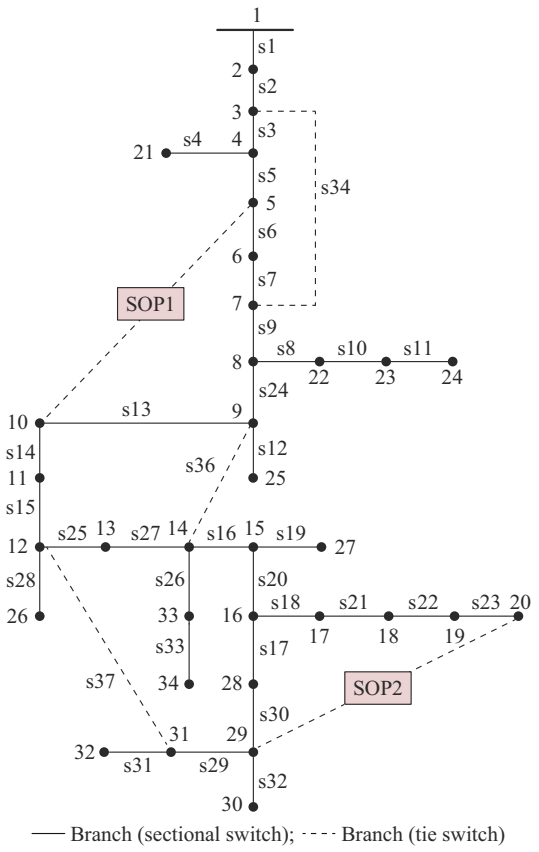


Fig. 2. Modified IEEE 34-bus system.

Then, $x_o^{\phi,j}\in[0,1]$ $(j=1,2,3,4)$ are used to represent the control variables of an SOP to ensure that the SOP control strategies satisfy constraints (2) and (3). The relationship between the floating points and the corresponding control strategies can be expressed as:

$$\begin{cases} P_{o,i}^{\mathrm{SOP},\phi}=x_o^{\phi,2}S_o^{\mathrm{SOP},\phi}(\mathrm{sgn}(x_o^{\phi,1}-0.5)-\mathrm{sgn}(0.5-x_o^{\phi,1})) \\ P_{o,j}^{\mathrm{SOP},\phi}=-P_{o,i}^{\mathrm{SOP},\phi} \\ Q_{o,i}^{\mathrm{SOP},\phi}=x_o^{\phi,3}\sqrt{(S_{o,i}^{\mathrm{SOP},\phi})^2-(P_{o,i}^{\mathrm{SOP},\phi})^2} \\ Q_{o,j}^{\mathrm{SOP},\phi}=x_o^{\phi,4}\sqrt{(S_{o,j}^{\mathrm{SOP},\phi})^2-(P_{o,j}^{\mathrm{SOP},\phi})^2} \end{cases} \tag{16}$$

According to the state set $\boldsymbol{S}^{\mathrm{SP}}$ and action set $\boldsymbol{A}^{\mathrm{SP}}$, the state transition probability of SP between $\boldsymbol{S}_t^{\mathrm{SP}}$ and $\boldsymbol{S}_{t+1}^{\mathrm{SP}}$ can be written as:

$$P_{\boldsymbol{S}_t^{\mathrm{SP}}\boldsymbol{S}_{t+1}^{\mathrm{SP}}}^{\boldsymbol{A}_t^{\mathrm{SP}}}=p(\boldsymbol{S}_{t+1}^{\mathrm{SP}}|\boldsymbol{S}_t^{\mathrm{SP}},\boldsymbol{A}_t^{\mathrm{SP}}) \tag{17}$$

The goal of MDP is to find a series of optimal strategies that can maximize the cumulative reward $G_t$, as shown in (18). Note that $R_t^{\mathrm{SP}}$ is the reward during period $t$, which is modeled based on the objective function (3).

$$G_t=R_t^{\mathrm{SP}}+\gamma R_{t+1}^{\mathrm{SP}}+\gamma^2 R_{t+2}^{\mathrm{SP}}+\cdots \tag{18}$$

To achieve the target of maximizing $G_t$ while minimizing the operation cost, the reward $R_t^{\mathrm{SP}}$ is set as the penalty divided by the operation cost.

$$R_t^{\mathrm{SP}}=R(\boldsymbol{S}_t^{\mathrm{SP}},\boldsymbol{A}_t^{\mathrm{SP}})=\frac{1-(1-\lambda^{\mathrm{SP}})\xi}{\sum\limits_{\phi}c_t^L\left(\sum\limits_{ij\in B}(I_{ij,t}^{\phi})^2r_{ij}^{\phi}\right)\Delta_t+c_s\Delta d_t} \tag{19}$$

If $\boldsymbol{A}_t^{\mathrm{SP}}$ satisfies the operation constraints (4) - (12), $\lambda^{\mathrm{SP}}=1$ and $R_t^{\mathrm{SP}}>0$; otherwise, $\lambda^{\mathrm{SP}}=0$ and $R_t^{\mathrm{SP}}<0$. For example, after action $\boldsymbol{A}_t^{\mathrm{SP}}$ generated by the agent is transmitted to the environment, it does not satisfy the voltage constraint of the system, then the environment gives the agent a negative reward, i. e., punishment. Conversely, when $\boldsymbol{A}_t^{\mathrm{SP}}$ satisfies the constraint, the environment will give the agent a positive reward. And the smaller the operation cost of $\boldsymbol{A}_t^{\mathrm{SP}}$, the greater the value of the $R_t^{\mathrm{SP}}$.

In the state $\boldsymbol{S}_t^{\mathrm{SP}}$ with action $\boldsymbol{A}_t^{\mathrm{SP}}$, the expectation of $G_t$ can be defined as state-action value $Q(\boldsymbol{S}_t^{\mathrm{SP}},\boldsymbol{A}_t^{\mathrm{SP}})$. It can be expressed in a recursive form called the Behrman equation.

$$Q(\boldsymbol{S}_t^{\mathrm{SP}},\boldsymbol{A}_t^{\mathrm{SP}})=E(G_t|\boldsymbol{S}_t^{\mathrm{SP}},\boldsymbol{A}_t^{\mathrm{SP}})=E(R(\boldsymbol{S}_t^{\mathrm{SP}},\boldsymbol{A}_t^{\mathrm{SP}}))+\gamma\sum_{\boldsymbol{S}_{t+1}^{\mathrm{SP}}\in\boldsymbol{S}^{\mathrm{SP}}}P_{\boldsymbol{S}_t^{\mathrm{SP}}\boldsymbol{S}_{t+1}^{\mathrm{SP}}}^{\boldsymbol{A}_t^{\mathrm{SP}}}Q(\boldsymbol{S}_{t+1}^{\mathrm{SP}},\boldsymbol{A}_{t+1}^{\mathrm{SP}}) \tag{20}$$

The method of solving SP-MDP is to find a set of optimal SP sequence control strategies to maximize the $Q$-value. The above process transforms an SP problem into an SP-MDP, whose brief framework is shown in Fig. 3.

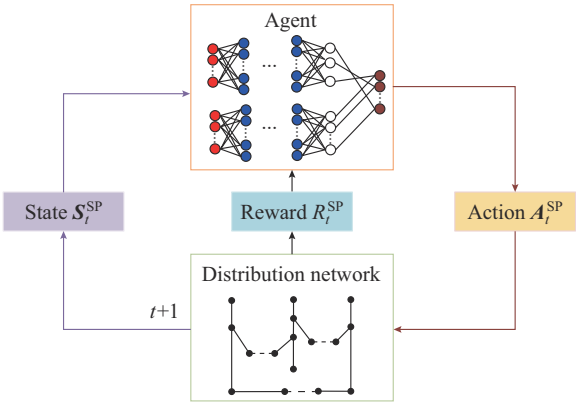


Fig. 3. MDP framework of sequential DN reconfiguration.

Under the framework of SP-MDP, the agent generates topology and SOP control action $\boldsymbol{A}_t^{\mathrm{SP}}$ according to the state $\boldsymbol{S}_t^{\mathrm{SP}}$ of the DN during time $t$. And $\boldsymbol{A}_t^{\mathrm{SP}}$ is transmitted to the DN for power flow calculation to get a reward $R_t^{\mathrm{SP}}$. Then, the above operations are repeated at next time step. Finally, a series of strategies that can maximize $G_t$ are learned through a closed-loop iteration.

However, it is worth noting that in the realistic DN, the change of bus injection power state between adjacent periods is an uncertain random process, which is affected by the weather and the user's electricity consumption behavior. Thus, it is difficult to give an explicit mathematical expression for the state transition probability in the SNR-MDP. Therefore, the model-free DRL algorithm with neural networks (NNs) is used to solve the SP-MDP model.

## III. DRL-BASED SP-MDP SOLVING METHOD

In this section, a DRL joint optimization solution method based on double deep $Q$ network (DDQN) and soft actor-critic (SAC) framework is constructed to exploit the advantages of different DRL methods for discrete and continuously variable control. Then, the BDDQN based on the fundamental loop matrix is proposed, converting the reconfiguration decision problem into a multi-dimensional action space decision-making problem. Finally, the MPSAC based on the multi-policy network is proposed to learn three-phase SOP control strategies.

### A. Solution Framework for SP-MDP

DRL combines deep learning and RL. The perception of deep learning is used to solve the modeling problems of policy and value function. And the decision-making ability of RL is used to define problems and optimize goals. The popular DRL algorithms for solving the MDP problem are DDQN that controls discrete variables and SAC that controls continuous variables. The detailed introduction of DDQN and SAC can refer to [30] and [31].

However, the state and action set will be too large due to many combinations of various control elements in the DN and the strong coupling. The agent cannot perform compelling exploration and training. Therefore, this paper proposes a solving method based on improved DDQN and SAC to realize the optimal joint control of DNR and SOP. In this paper, the proposed method is divided into two stages: offline training and online execution. The joint optimization framework of DRL is shown in Fig. 4.

In the offline training stage, BDDQN and MPSAC (BD-AC) agents learn the DN topology and SOP control strategy. Two agents share the reward $R_t^{\mathrm{SP}}$ and cooperate to learn the SP control strategies that maximize the cumulative reward. In the online execution stage, the DRL-based method can make decisions directly according to the real-time DN measurement data [32] to realize the optimal control of SP.

### B. Proposed BDDQN Approach

The DDQN uses two NNs, i.e., $Q$ network $\boldsymbol{Q}_{\mathrm{pre}}$ and target $Q$ network $\boldsymbol{Q}_{\mathrm{tar}}$, to approximate state-action value (20) with the same architecture. Assuming that the data $\{\boldsymbol{S}_t^{\mathrm{SP}},\boldsymbol{A}_t^{\mathrm{SNR}},R_t^{\mathrm{SP}},\boldsymbol{S}_{t+1}^{\mathrm{SP}}\}$ are sampled from experience pool $M^{\mathrm{SNR}}$, $\boldsymbol{S}_{t+1}^{\mathrm{SP}}$ is input to the $Q$ network, and the action $\hat{\boldsymbol{A}}_{t+1}^{\mathrm{SNR}}$ can be selected based on the greedy strategy.

$$\hat{\boldsymbol{A}}_{t+1}^{\mathrm{SNR}}=\arg\max\boldsymbol{Q}_{\mathrm{pre}}(\boldsymbol{S}_{t+1}^{\mathrm{SP}},\sim;\beta)\quad \varsigma\le\varepsilon \tag{21}$$

where $\varsigma\in[0,1]$ is a random number; $\varepsilon$ is the greedy selection factor. If $\varsigma>\varepsilon$, $\hat{\boldsymbol{A}}_{t+1}^{\mathrm{SNR}}$ is a random action. The action with the largest $Q$-value is the SNR strategy that can minimize the operation cost.

However, applying DDQN to DNR tasks requires addressing the combined growth of the number of possible actions and the number of action dimensions [31]. Due to the switch state in the DN being restricted by its operation characteristics, the one-dimensional action space cannot be used to visually describe the action relationship between switches. Taking Fig. 2 as an example, the IEEE 34-bus system has $6\times5\times8=240$ switch combinations. DDQN needs to select an optimal strategy from 240 strategies at each iteration. Moreover, when the scale of the DN expands, the possible switch combinations are shown explosive growth, which significantly increases the difficulty in learning the SNR control strategy.

To solve this problem, we propose a BDDQN based on branching dueling $Q$-network (BDQ) [33].

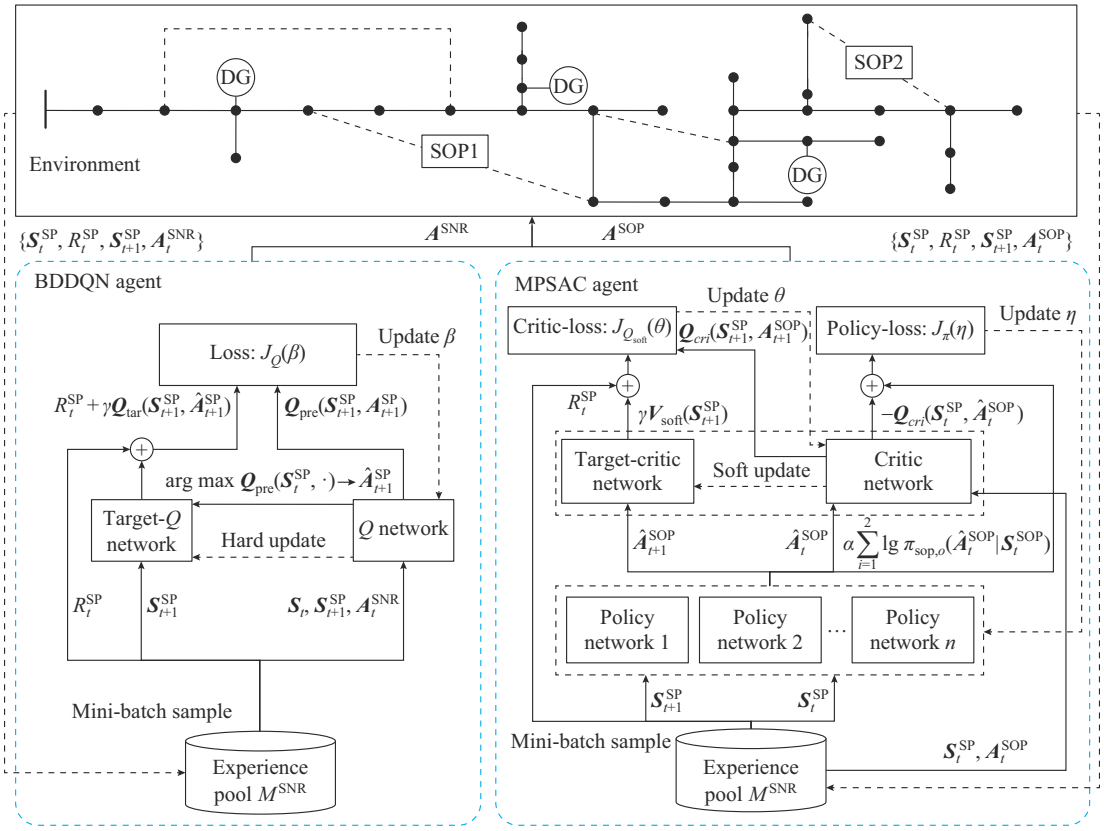


Fig. 4. Joint optimization framework of DRL.

The BDQ has the same number of sub-actions for each action dimension. However, the number of switches in each loop usually differs in the DN. Therefore, compared with BDQ, the significant advantage of BDDQN is that the length of the $Q$-value vector in each dimension can be adjusted adaptively according to the number of switches in each loop.

In Fig. 5, the improved $Q$ network can output the potential action value $B_d(\boldsymbol{S}_t^{SP}, A_{t,l,h}^{SNR})$ of each switch according to $\boldsymbol{S}_t^{SP}$.

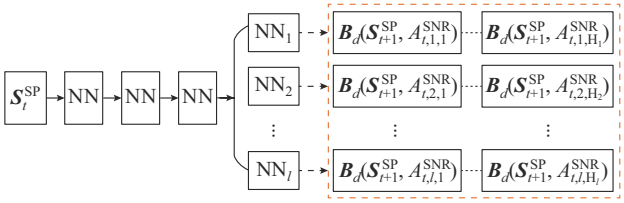


Fig. 5. Structure of $Q$ network.

Moreover, the mean operator [33] is used to express the $Q$-value matrix:

$$\boldsymbol{Q}_{pre}(\boldsymbol{S}_t^{SP}, A_{t,l,h}^{SNR}) = \boldsymbol{B}_d(\boldsymbol{S}_t^{SP}, A_{t,l,h}^{SNR}) - \frac{1}{Z}\sum_{l,h}\boldsymbol{B}_d(\boldsymbol{S}_t^{SP}, A_{t,l,h}^{SNR}) \quad (22)$$

where $l \in \{1, 2, \ldots, L\}$; and $h \in \{1, 2, \ldots, H_l\}$. Therefore, the element in $\boldsymbol{Q}_{pre}$ is the value produced by the switch action in the state $\boldsymbol{S}_t^{SP}$. For example, $\boldsymbol{Q}_{pre}(\boldsymbol{S}_t^{SP}, A_{t,1,1}^{SNR})$ is the action value of branch 16 that disconnects in the IEEE 34-bus system.

Then, the reconfiguration result in $t+1$ for each loop can be selected according to greedy selection.

$$\boldsymbol{A}_{t+1,l}^{SNR} = \underset{A_{t+1,l,h}^{SNR}}{\arg\max}\,(\boldsymbol{Q}_{pre}(\boldsymbol{S}_{t+1}^{SP}, A_{t+1,l,h}^{SNR}), \beta) \quad \varsigma \le \varepsilon \quad (23)$$

According to (21), BDDQN can select a switch with the maximum value in each row of $\boldsymbol{Q}_{pre}(\boldsymbol{S}_{t+1}^{SP}, A_{t+1,l,h}^{SNR})$ to constitute a complete reconfiguration strategy. In this way, the original one-dimensional complex decision-making process can be transformed into a multi-dimensional simple decision-making process.

The target $Q$ network is used to evaluate the $Q$-value of the SP strategy given by the agent. It can be expressed as:

$$\boldsymbol{Q}_{tar}(\boldsymbol{S}_t^{SP}, \boldsymbol{A}_{t,l}^{SNR}) = \begin{cases} R_t^{SP} & t = |T_{SP}| \\ R_t^{SP} + \gamma \boldsymbol{Q}_{tar}(\boldsymbol{S}_{t+1}^{SP}, \underset{\boldsymbol{A}_{t+1,l}^{SNR}}{\arg\max}\,\boldsymbol{Q}_{pre}(\boldsymbol{S}_{t+1}^{SP}, A_{t+1,l}^{SNR}; \beta); \hat{\beta}) & t < |T_{SP}| \end{cases} \quad (24)$$

Then, the $Q$ target-value and $Q$-value are input to the loss function $J_Q(\beta)$ to update the $Q$ network parameters.

$$J_Q(\beta) = \frac{1}{|D|}\sum_{\{\boldsymbol{S}_t^{SP}, \boldsymbol{A}_t^{SNR}, R_t^{SP}, \boldsymbol{S}_{t+1}^{SP}\} \in M^{SNR}} \frac{1}{L}\sum_{l=1}^{L}(\boldsymbol{Q}_{tar}(\boldsymbol{S}_t^{SP}, \boldsymbol{A}_{t,l}^{SNR}; \hat{\beta}) - \boldsymbol{Q}_{pre}(\boldsymbol{S}_t^{SP}, \boldsymbol{A}_{t,l}^{SNR}; \beta))^2 \quad (25)$$

BDDQN agent learns the optimal SNR control strategy by adjusting its $Q$ network parameters $\hat{\beta}$ and $\beta$ towards minimizing the operation cost objective.

### C. Proposed MPSAC Approach

As shown in Fig. 4, MPSAC has multiple policy networks, each policy network $\pi_{sop,o}$ outputs the three-phase

control variables of a single SOP and shares a critic network. The critic network and target critic network have the same structure.

Unlike BDDQN, MPSAC algorithm learns policy networks and critic networks. The critic network evaluates SOP control actions generated by the policy networks to minimize the operation cost. Moreover, to improve the exploration efficiency of the algorithm, the optimization objective of the MPSAC is to maximize the sum of cumulative return and action entropy.

$$J(\pi_{\text{sop}})=\max\sum_{t=1}^{|T_{\text{SP}}|}E\left(R_t^{\text{SP}}+\alpha\sum_{o=1}^{O}H_o\left(\pi_{\text{sop},o}(\sim|\boldsymbol{S}_t^{\text{SP}})\right)\right) \tag{26}$$

where $H_o(\cdot)$ is the action-entropy function of $\pi_{\text{sop},o}$; $\pi_{\text{sop},o}(\sim|\boldsymbol{S}_t^{\text{SP}})$ generates a distribution of SOP control strategies in $\boldsymbol{S}_t^{\text{SP}}$; and $\hat{\boldsymbol{A}}_t^{\text{SOP}}$ is sampled from the distribution. The larger the value of $H_o(\cdot)$, the more random the SOP control action generated by the policy network. $H_o(\cdot)$ can be expressed as [31]:

$$H_o\left(\pi_{\text{sop},o}(\sim|\boldsymbol{S}_t^{\text{SP}})\right)=\mathbb{E}_{\hat{\boldsymbol{A}}_{t+1}^{\text{SOP}}\sim\pi_{\text{sop},o}(\cdot|\boldsymbol{S}_t^{\text{SP}})}\left[-\lg\pi_{\text{sop},o}\left(\hat{\boldsymbol{A}}_t^{\text{SOP}}|\boldsymbol{S}_t^{\text{SP}}\right)\right] \tag{27}$$

Assume that the SOP control action is Gaussian distributed with mean $\mu(\boldsymbol{S}_t^{\text{SP}})$ and covariance $\sigma^2(\boldsymbol{S}_t^{\text{SP}})$, where $\mu(\boldsymbol{S}_t^{\text{SP}})$ and $\sigma^2(\boldsymbol{S}_t^{\text{SP}})$ are parameterized by the policy network [31]. The policy function can be expressed as:

$$\pi_{\text{sop},o}(\sim|\boldsymbol{S}_t^{\text{SP}})=\mathbb{N}(\boldsymbol{\mu}_o(\boldsymbol{S}_t^{\text{SP}}),\boldsymbol{\sigma}_o^2(\boldsymbol{S}_t^{\text{SP}})) \tag{28}$$

MPSAC is divided into two parts: policy evaluation and policy improvement. Assuming that the data $\{\boldsymbol{S}_t^{\text{SP}},\boldsymbol{A}_t^{\text{SOP}},R_t^{\text{SP}},\boldsymbol{S}_{t+1}^{\text{SP}}\}$ are sampled from experience pool $M^{\text{SOP}}$, in the policy evaluation part, the action value of $\mu(\boldsymbol{S}_t^{\text{SP}})$ is evaluated through the target critic network, which can be expressed as:

$$\begin{cases}\boldsymbol{Q}_{\text{soft}}(\boldsymbol{S}_t^{\text{SP}},\hat{\boldsymbol{A}}_t^{\text{SOP}})=R_t^{\text{SP}}+\gamma\boldsymbol{V}_{\text{soft}}(\boldsymbol{S}_{t+1}^{\text{SP}})\\ \boldsymbol{V}_{\text{soft}}(\boldsymbol{S}_{t+1}^{\text{SP}})=\boldsymbol{Q}_{cri,t}(\boldsymbol{S}_{t+1}^{\text{SP}},\hat{\boldsymbol{A}}_{t+1}^{\text{SOP}})+\alpha\sum_{o=1}^{O}H_o(\pi_{\text{sop},o}(\sim|\boldsymbol{S}_{t+1}^{\text{SP}}))\end{cases} \tag{29}$$

where $\hat{\boldsymbol{A}}_t^{\text{SOP}}$ is equivalent to $\mu(\boldsymbol{S}_t^{\text{SP}})$. Then, to learn the parameters $\theta$ and $\hat{\theta}$ of the critic networks, $\boldsymbol{Q}_{\text{soft}}(\boldsymbol{S}_t^{\text{SP}},\hat{\boldsymbol{A}}_t^{\text{SOP}})$ and $\boldsymbol{Q}_{cri}(\boldsymbol{S}_t^{\text{SP}},\hat{\boldsymbol{A}}_t^{\text{SOP}};\theta)$ are input to the critic loss function $J_{Q_{\text{soft}}}(\theta)$, which can be expressed as:

$$J_{Q_{\text{soft}}}(\theta)=\frac{1}{|D|}\sum_{\{\boldsymbol{S}_t^{\text{SP}},\boldsymbol{A}_t^{\text{SOP}},R_t^{\text{SP}},\boldsymbol{S}_{t+1}^{\text{SP}}\}\in M^{\text{SOP}}}(\boldsymbol{Q}_{cri}(\boldsymbol{S}_t^{\text{SP}},\hat{\boldsymbol{A}}_t^{\text{SOP}};\theta)-\boldsymbol{Q}_{\text{soft}}(\boldsymbol{S}_t^{\text{SP}},\hat{\boldsymbol{A}}_t^{\text{SOP}};\hat{\theta}))^2 \tag{30}$$

In the policy improvement part, the optimization of policy network parameters $\eta$ is achieved by minimizing the policy loss function $J_\pi(\eta)$, which can be expressed as [31]:

$$J_\pi(\eta)=\frac{-1}{|D|}\sum_{\{\boldsymbol{S}_t^{\text{SP}},\boldsymbol{A}_t^{\text{SOP}},R_t^{\text{SP}},\boldsymbol{S}_{t+1}^{\text{SP}}\}\in M^{\text{SOP}}}\alpha\sum_{o=1}^{O}H_o(\pi_{\text{sop},o}(\sim|\boldsymbol{S}_t^{\text{SP}};\eta))+\boldsymbol{Q}_{cri}(\boldsymbol{S}_t^{\text{SP}},\hat{\boldsymbol{A}}_t^{\text{SOP}};\theta) \tag{31}$$

MPSAC agent can also learn the optimal SOP control strategy by adjusting its NN parameters $\eta$, $\theta$, and $\hat{\theta}$.

### *D. Summary of Proposed Algorithm*

The above process can establish the BD-AC algorithm in multi-dimensional action space. Furthermore, the optimal SP strategy to reduce operation cost can be found by iterative training NN. The specific flow is summarized in Algorithm 1.

**Algorithm 1**: DRL-based SP control algorithm

**1) Offline training**
**Input**: historical dataset, discount factor $\gamma$, batch number $|D|$, action space dimension
**Initialize** experience pool $M^{\text{SNR}}$ and $M^{\text{SOP}}$, parameters $\beta$, $\theta$, and $\eta$
**For** each episode, **do**
  **Initialize** sequence $\boldsymbol{S}_t^{\text{SP}}(t=1)$
  **For** decision time step $t\in T_{\text{SP}}$, **do**
    **If** $\varsigma>\varepsilon$, **then**
      Select a random action $\boldsymbol{A}_t^{\text{SNR}}$ and $\boldsymbol{A}_t^{\text{SOP}}$
    **Else**
      $\boldsymbol{A}_t^{\text{SNR}}=\arg\max_{\boldsymbol{A}_{t,l,h}^{\text{SNR}}}(\boldsymbol{Q}_{\text{pre}}(\boldsymbol{S}_t^{\text{SP}},\sim;\beta))$
      $\boldsymbol{A}_t^{\text{SOP}}=\mathbb{E}(\hat{\boldsymbol{A}}_{t+1}^{\text{SOP}}\sim\pi_{\text{sop},o}(\sim|\boldsymbol{S}_t^{\text{SP}}))$
    **End if**
    Calculate the reward $R_t^{\text{SP}}$ according to (19)
    Index $\boldsymbol{S}_{t+1}^{\text{SP}}$ from the historical dataset of bus injection power
    Store $\boldsymbol{S}_t^{\text{SP}},\boldsymbol{A}_t^{\text{SNR}},R_t^{\text{SP}},\boldsymbol{S}_{t+1}^{\text{SP}}$ and $\boldsymbol{S}_t^{\text{SP}},\boldsymbol{A}_t^{\text{SOP}},R_t^{\text{SP}},\boldsymbol{S}_{t+1}^{\text{SP}}$ in $M^{\text{SNR}}$ and $M^{\text{SOP}}$, respectively
    Set $\boldsymbol{S}_t^{\text{SP}}=\boldsymbol{S}_{t+1}^{\text{SP}}$
    **If** $|M^{\text{SNR}}|$ and $|M^{\text{SOP}}|>|D|$, **then**
    Sample $|D|$ pairs of $\{\boldsymbol{S}_j^{\text{SP}},\boldsymbol{A}_j^{\text{SNR}},R_j^{\text{SP}},\boldsymbol{S}_{j+1}^{\text{SP}}\}$ from $M^{\text{SNR}}$
    Calculate $\boldsymbol{Q}_{\text{pre}}$ and $\boldsymbol{Q}_{\text{tar}}$ by (22) and (24)
    Use (25) to calculate $Q$ network loss
    Update all the parameters $\beta$ using the Adam [30]
    Every $C$ step rest $\hat{\beta}\leftarrow\beta$
    Sample $|D|$ pairs of $\{\boldsymbol{S}_j^{\text{SP}},\boldsymbol{A}_j^{\text{SOP}},R_j^{\text{SP}},\boldsymbol{S}_{j+1}^{\text{SP}}\}$ from $M^{\text{SOP}}$
    Calculate $\boldsymbol{Q}_{\text{soft}}$ by (29)
    Use (30) to calculate critic network loss
    Update the parameter $\theta$ using the Adam
    Use (31) to calculate policy network loss
    Update the parameter $\eta$ using the Adam
    Every $C$ step rest $\hat{\theta}\leftarrow\tau\theta+(1-\tau)\hat{\theta}$
    **End if**
  **End For**
**End for**
**2) Online execution**
**For each day do**
  **For** $t=1:24$ **do**
    Collect real-time bus injection power $\boldsymbol{S}_t^{\text{SP}}$ in $t$
    Output $\boldsymbol{Q}_{\text{pre}}(\boldsymbol{S}_t^{\text{SP}},\sim)$ and $\pi_{\text{sop},o}(\sim|\boldsymbol{S}_t^{\text{SP}})$
    Use (16) and $\arg\max(\boldsymbol{Q}_{\text{pre}})$ to generate the decision result
    Execute SOP and topology control strategy
  **End for**
**End for**

In summary, the difference between the proposed method and the traditional SNR method is as follows.

The traditional SNR method requires an optimization algorithm to find an offline solution to optimize the objective function value. Moreover, the traditional SP method uses the predicted value of load and DG output to obtain the SP solution, that is why it requires considering the uncertainty of DG output and load.

However, after the BD-AC agent is trained off-line, the proposed method has learned the mapping relationship $\pi(\boldsymbol{S}^{\text{SP}},\boldsymbol{A}^{\text{SP}})$ from the historical data. And in the online execution stage, the real-time bus injection power collected by the SCADA system or PMU system [19] can be input into the trained agent to obtain SP strategy immediately. Therefore,

the proposed method can reduce the influence of DG and load uncertainties, and immediately resolve the problem of overvoltage caused by high PV permeability.

## IV. Case Study

In this section, to verify the performance of the proposed SP-MDP method, comprehensive case studies on IEEE standard test systems are conducted. The experimental data and the algorithm setup are first presented. Then, the IEEE 34-bus system is used to verify the superiority of BD-AC. Furthermore, the IEEE 123-bus system [34] with high PV penetration is used to verify if BD-AC can avoid overvoltage problem while optimizing operation cost.

### A. Experimental Data and Algorithm Setup

In this paper, load and DG data from 2012 [35] and 2014 [36] Global Energy Forecasting Competition are used to generate plenty of load and DG profiles. Then, the resultant load and DG instances are normalized to match the scale of power demands in the simulated system to train the BD-AC agent. Next, the proposed method is trained using "Pytorch" on the NVIDIA RTX 3090 GPU with 12 GB RAM. And the environment of SP-MDP offline training is established through the software of OpenDSS. $c_t^L$ is 0.16 \$/kWh and $c_s$ is \$2 [1]. The hyperparameters of the different methods are provided in Table I. The number of hidden layers for all NNs is three.

TABLE I
Hyperparameters of Different Methods

| Method | Hyperparameter | IEEE 34-bus system | IEEE 123-bus system |
|---|---|---|---|
| BDDQN | Minibatch size | 32 | 128 |
| | Discount factor | 0.99 | 0.99 |
| | Learning rate | $3\times10^{-4}$ | $8\times10^{-4}$ |
| | Number of hidden units | 64 | 128 |
| MPSAC | Minibatch size | 128 | 256 |
| | Discount factor | 0.99 | 0.99 |
| | Learning rate | $3\times10^{-4}$ | $3\times10^{-4}$ |
| | Number of hidden units | 128 | 256 |

### B. IEEE 34-bus System

The IEEE 34-bus system is shown in Fig. 2. The SOPs are installed on branch S37 and branch S38 to replace tie switches. The capacity of each SOP is 100 kVA. The location and capacity of DGs are shown in Table II. The algorithm in [37] is used to obtain the static DNR and SOP control results.

TABLE II
Location and Capacity of DGs

| Bus number | Capacity (kVA) | | | Power factor | DG type |
|---|---|---|---|---|---|
| | Phase A | Phase B | Phase C | | |
| 5 | 250 | 250 | 250 | 0.90 | Wind power generation |
| 18 | 250 | 250 | 0 | 0.95 | Solar power generation |
| 22 | 200 | 0 | 0 | 0.90 | Wind power generation |

During the training process of the BD-AC, we record the weights of the NN every 50 epochs, which are used to evaluate the performance of the method on the testing data. The cumulative operation cost for different methods is shown in Fig. 6. The mixed-integer linear programming (MILP) combines the dynamic DNR [38] and the polyhedral-based approximation method [39].

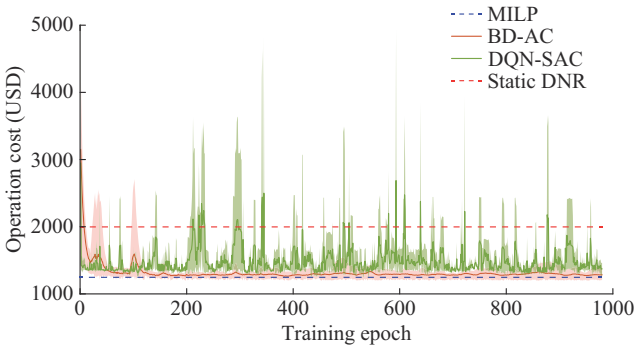


Fig. 6. Cumulative operation cost for different methods.

In Fig. 6, the DQN-SAC method cannot achieve optimal decisions within the prescribed training steps. However, the multi-dimensional $Q$ network and multi-policy network of BD-AC agent reduce the action space of optimization decisions to improve the search efficiency, decreasing the operation costs by 5.6% compared with DQN-SAC. Moreover, the operation cost of the MILP is \$1100.8, which is only 0.9% lower than that of the proposed method. Therefore, it can be concluded that the proposed method can converge to the optimal decision faster and more stable.

Then, two scenarios are compared in detail to further illustrate the superiority of the proposed method.

Scenario 1: the unbalanced optimal operation with static DNR and SOP control.

Scenario 2: the unbalanced optimal operation with SP.

In Fig. 7, the total costs of scenarios 1 and 2 are \$1491.5 and \$1209.7, respectively. This is because the static DNR frequently controls switches to minimize energy loss. The total energy losses of scenarios 1 and 2 are $8.26\times10^3$ kWh and $9.22\times10^3$ kWh, respectively. After considering the switch action cost, the total operation cost of scenario 2 is 19.58% less than that of scenario 1. Therefore, it can be concluded that our proposed method can effectively reduce the operation cost compared with static DNR with SOP. The action value matrix of switch at the 20$^{th}$ hour is shown in Fig. 8.

It can be observed from Fig. 8 that the BDDQN agent can output the action value of all switches. The maximum value of each row is 1.566, 0.8759, and 1.266, respectively. According to this value matrix, the switch with the maximum value in each loop can be selected to compose a reconfiguration strategy. Thus, the optimal topology state is disconnected with branches s7, s25, and s30, which can reduce the power loss by 12.24% compared with the initial state. It is worth noting that the values in the matrix represent the estimated value of the cumulative reward for each switch action, but not the system operation cost.

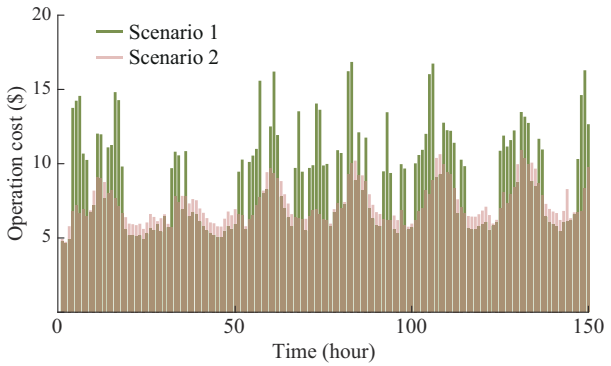


Fig. 7. Operation cost of different scenarios.

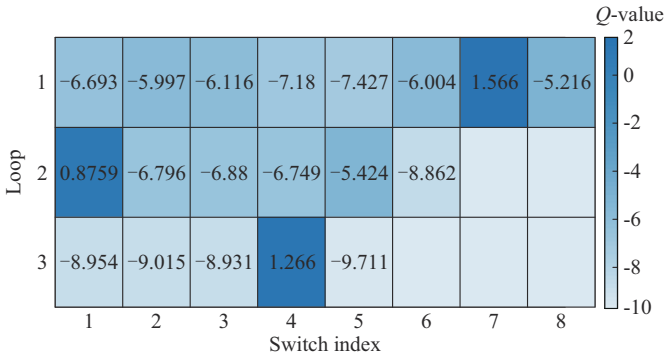


Fig. 8. Action value matrix of switch.

Moreover, to verify the adaptation of the proposed method against the load power mutation, the following three cases based on the testing week data are considered.

Case 1: the load demand and DG output are reduced by 15% and increased by 15%, respectively.

Case 2: the DG output is increased by 15%.

Case 3: the load demand and DG output are increased by 15% and reduced by 15%, respectively.

It can be observed from Table III that the proposed method can still make ideal decisions to reduce the operation cost even when the bus injected power changes significantly compared with the historical data. Finally, the proposed method is compared with the model-based method. The result is provided in Table IV.

TABLE III
COMPARISON OF DIFFERENT CASES FOR IEEE 34-BUS SYSTEM

| Case | Energy loss (kWh) | Operation cost ($) | The maximal bus voltage deviation (p.u.) | Operation cost reduction (%) |
|---|---|---|---|---|
| Case 1 | $8.05\times10^4$ | $1.04\times10^3$ | 0.054 | 15.18 |
| Case 2 | $9.21\times10^4$ | $1.20\times10^3$ | 0.059 | 15.39 |
| Case 3 | $1.12\times10^4$ | $1.45\times10^3$ | 0.064 | 16.01 |

It can be observed from Table IV that the static DNR with SOP can minimize energy loss, which is 40 kW, 15 kW, and 40 kW less than BD-AC, MILP, and HFWA (which is an improved firework algorithm based on heuristic rules). However, the SP result of BD-AC only disconnects branches 7, 25, and 30. Therefore, comprehensively considering the total of power loss cost and switch action cost, the proposed method reduces the operation cost compared with HFWA. Moreover, MILP can obtain the optimal SP strategy, and the operation cost reduction ratio is 0.92% higher than that of BD-AC. Therefore, it can be concluded that the proposed method can solve the SP optimization problem effectively.

TABLE IV
COMPARISON OF DIFFERENT METHODS FOR IEEE 34-BUS SYSTEM

| Method | Time and corresponding open switch | Switch action number | Energy loss (kWh) | Operation cost ($) | Operation cost reduction (%) |
|---|---|---|---|---|---|
| Static DNR with SOP | 00:00-06:00: 7, 25, 30<br>07:00-09:00: 7, 25, 17<br>10:00-14:00: 7, 27, 17<br>15:00-16:00: 6, 25, 17<br>17:00-18:00: 7, 25, 17<br>19:00:23:00: 7, 25, 30 | 24 | $1.028\times10^3$ | 213.14 | |
| BD-AC | 00:00-23:00: 7, 25, 30 | 6 | $1.071\times10^3$ | 183.21 | 15.28 |
| MILP | 01:00-09:00: 7, 25, 30<br>10:00-24:00: 7, 25, 17 | 8 | $1.042\times10^3$ | 181.97 | 16.20 |
| HFWA [37] | 01:00-09:00: 7, 27, 30<br>10:00-15:00: 7, 27, 17<br>16:00-24:00: 7, 13, 30 | 12 | $1.035\times10^3$ | 190.42 | 13.38 |

To illustrate the superiority of DRL-based method in computing efficiency, the comparison of computational efficiency of different methods is shown in Table V. Note that the training time is the CPU time of 1000-episode training process for the DRL algorithms, and the testing time is the CPU time for solving the SP problem in Table IV.

TABLE V
COMPARISON OF COMPUTATIONAL EFFICIENCY OF DIFFERENT METHODS

| Type | Method | Value |
|---|---|---|
| Training time | DQN-SAC | 10.36 hours |
| | BD-AC | 12.57 hours |
| Testing time | DQN-SAC | 2.41 ms |
| | BD-AC | 2.60 ms |
| | HFWA | 171.81 s |
| | MILP | 251.57 s |

In terms of training time, the proposed BD-AC increases the branching structure of the $Q$ network and has multiple strategic networks. Therefore, the training time is longer than the traditional DQN-SAC algorithm. However, as shown in Fig. 6, the performance of the BD-AC is better than the traditional method, and the training processes of the traditional method converge slowly. Besides, although the

calculation time of HFWA and MILP is only 171.81 s and 251.57 s, respectively, it relies on the day-ahead forecast data of DG and load. So the uncertainty needs to be considered. Our proposed method can make the control decision through the trained BD-AC agent without complex calculation. Thus, the decision time of a single period is only 1.76 ms, which can ensure that the agent makes decision according to the real-time load and DG data, thereby reducing the impact of system uncertainty on the optimization result.

According to the above results, it can be concluded that the proposed method can deal with the SP problem efficiently and accurately in the IEEE 34-bus system.

### *C. Modified IEEE 123-bus System with High Penetration of PVs*

In this subsection, the modified IEEE 123-bus system in Fig. 9 is used to verify the decision-making ability of the proposed method in a more complex system. Ten single-phase PVs with a constant power factor 0.9 are integrated into the network. The parameters of PVs are shown in Table VI. The capacity of each SOP is 500 kVA.

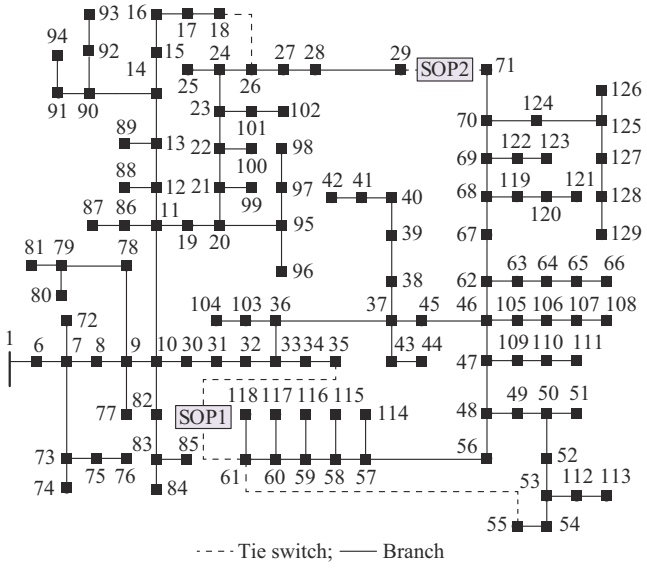


Fig. 9. Modified IEEE 123-bus system.

TABLE VI
PARAMETERS OF PVS

| Location | Phase | Capacity (kVA) | Location | Phase | Capacity (kVA) |
|---|---|---|---|---|---|
| 15 | A | 355 | 117 | B | 266 |
| 99 | C | 355 | 118 | B | 533 |
| 111 | C | 533 | 65 | C | 355 |
| 113 | C | 266 | 123 | B | 266 |
| 57 | B | 533 | 126 | A | 266 |

In this subsection, four scenarios are compared in detail.

Scenario 1: the initial operation state without optimization.

Scenario 2: the unbalanced optimal operation with DNR.

Scenario 3: the unbalanced optimal operation with static DNR and SOP control.

Scenario 4: the unbalanced optimal operation with SP.

We compare the optimization results of different scenarios in detail on the testing week, as shown in Table VII. Note that the overvoltage rate is the proportion of the amount of overvoltage in the testing week.

TABLE VII
OPTIMIZATION RESULTS OF DIFFERENT SCENARIOS

| Scenario | Energy loss (kW) | Switch action number | Operation cost ($) | Overvoltage rate (%) |
|---|---|---|---|---|
| 1 | $3.22\times10^4$ | 0 | $4.18\times10^3$ | 15.48 |
| 2 | $2.74\times10^4$ | 478 | $4.58\times10^3$ | 6.55 |
| 3 | $2.17\times10^4$ | 470 | $3.71\times10^3$ | 0 |
| 4 | $2.28\times10^4$ | 32 | $3.03\times10^3$ | 0 |

It can be observed from Table VII that scenarios 1 and 2 have serious system overvoltage problems. Scenarios 3 and 4 avoid the overvoltage issues by optimizing the control strategy for the SOP to provide additional active support to the system. However, scenario 3 leaves the economy out of consideration, and its switch action is too frequent. As a result in that its operation cost is significantly higher than that of scenario 4. Scenario 4 can reduce the operation cost by 18.72% compared with scenario 1 through minimizing the number of switch action. Therefore, it can be concluded that our proposed method also effectively reduces the operation cost in a complex unbalanced system.

Then, we select the day with the highest penetration of PVs in historical data to verify the ability of BD-AC to alleviate system overvoltage. The daily operation curves of three-phase total load and PV power are shown in Fig. 10. The maximum bus voltage deviation for different scenarios is shown in Fig. 11.

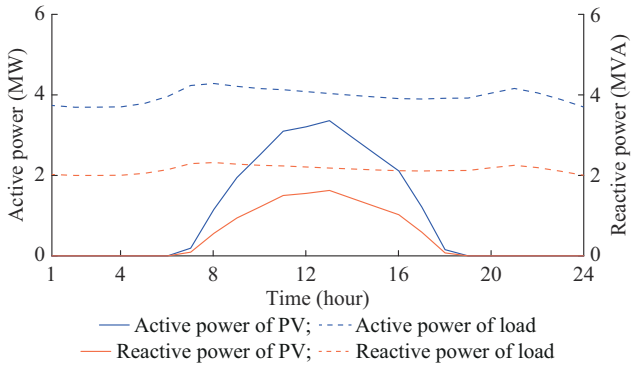


Fig. 10. Daily operation curves of three-phase total load and PV power.

As shown in Figs. 10 and 11, the range of active power penetration of PV at the 10th to 15th hour is 57%-83%, which leads to serious overvoltage problems in the DN. Although DNR reduces the maximum voltage deviation, it is still beyond the safe operation range. However, scenarios 3 and 4 can resolve the system overvoltage problem through DR and SOP joint optimization. To illustrate the effectiveness of the proposed method, the $Q$-value output of the switch action by BDDQN and the SOP action distribution output by MPSAC are shown in Figs. 12 and 13, respectively.

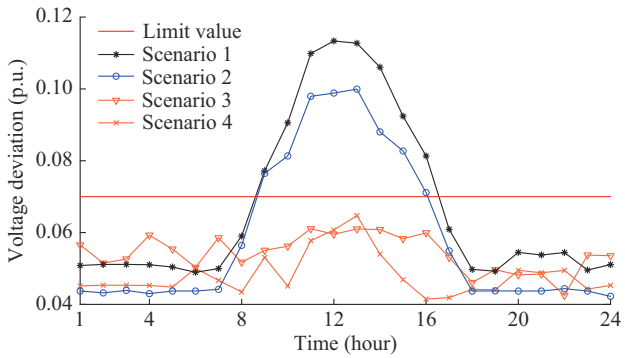


Fig. 11. The maximum bus voltage deviation for different scenarios.

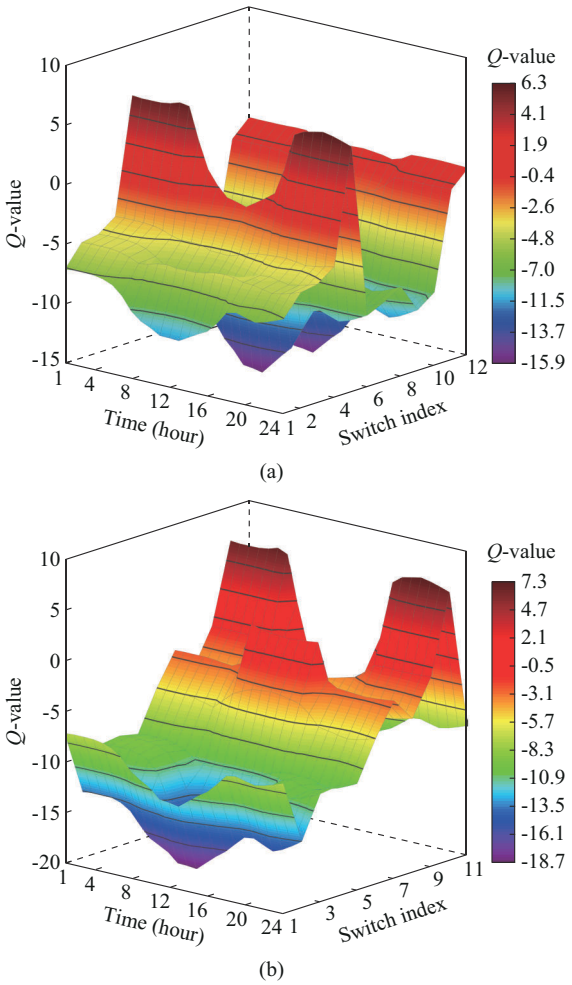


Fig. 12. *Q*-value outpout of switch action for test day. (a) Loop 1. (b) Loop 2.

In Fig. 13, *B* to *M* are the four decision variables of phases A, B, and C of SOP $x_o^{\phi,1}$,$x_o^{\phi,2}$,$x_o^{\phi,3}$, and $x_o^{\phi,4}$.

As shown in Figs.12 and 13, the pre-trained BD-AC agent directly outputs the switching action *Q*-value and SOP action distribution during each decision period according to the three-phase bus injection power of the DN. In Fig. 12, the proposed method selects the switch with the maximum *Q*-value in each loop to form an SNR strategy. Thus, the SNR strategy is that branches 18-26 and 58-59 are disconnected in the 10th to 16th hour, and branches 17-18 and 48-56 are disconnected during other periods. In Fig. 13, the actions of SOP follow Gaussian distribution. The proposed method takes the expected value of the distribution as the action value of SOP and then calculates the active and reactive power transmitted by SOP according to (16). Furthermore, the three-phase power transmission for SOP in scenario 4 is shown in Fig. 14.

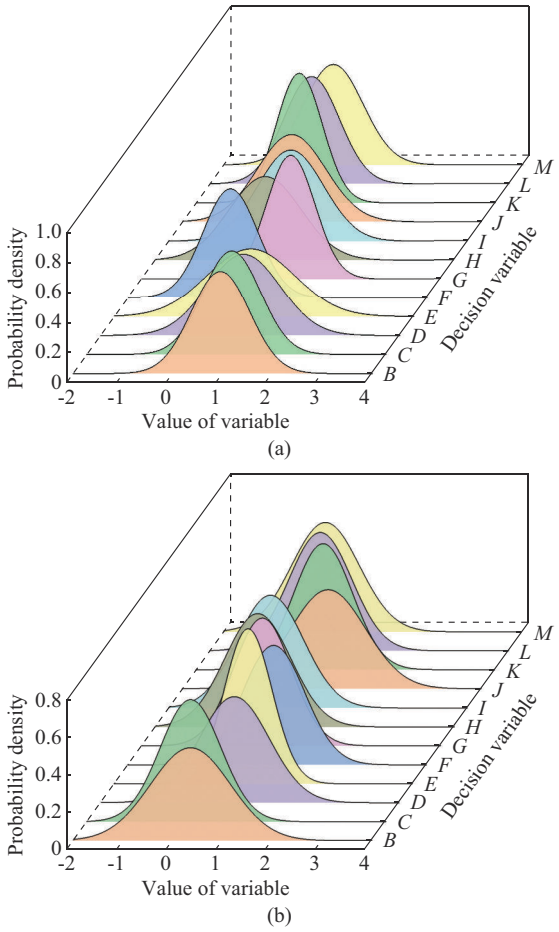


Fig. 13. Action distribution of SOP in the 13th hour. (a) SOP1. (b) SOP2.

As shown in Fig. 14, the BD-AC agent can adjust the three-phase control strategy of the SOP based on the system operation state, thus effectively providing active and reactive power supports to the system to mitigate overvoltage problems caused by PV. In addition, the network loss and operation cost of different scenarios on test day are shown in Fig. 15.

It can be observed from Fig. 15(a) that the network loss can be significantly reduced by adding SOP. In Fig. 15(b), the total operation cost of scenario 4 is \$349.1, which is 32.2%, 29.3%, and 25.1% lower than scenarios 1, 2, and 3, respectively. At the 10th hour, the operation cost of scenario 4 suddenly increases because the PV penetration is too high during this period, and the network topology needs to be reconfigured to reduce voltage deviation. Thus, the network structure changes from the disconnection of branches 17-18 and 48-56 to the disconnection of branches 18-26 and 58-59. At the 16th hour, the system will not have overvoltage risk due to the reduction of PV output power.

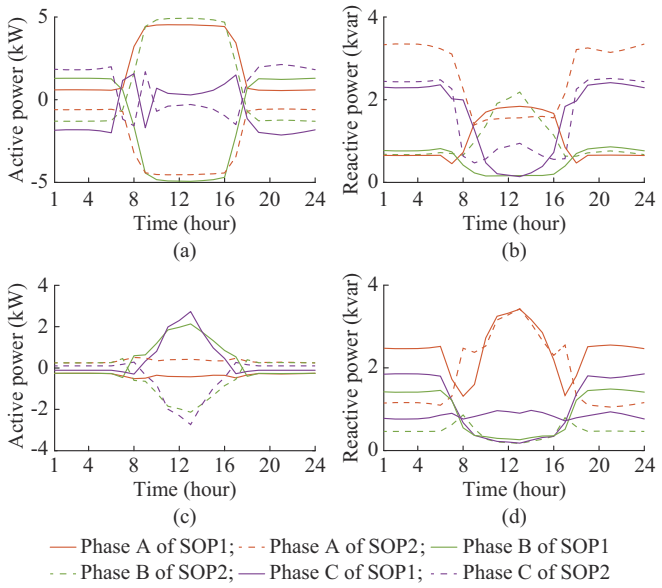


Fig. 14. Three-phase power transmission for SOP in scenario 4. (a) Active power of SOP1. (b) Reactive power of SOP1. (c) Active power of SOP2. (d) Reactive power of SOP2.

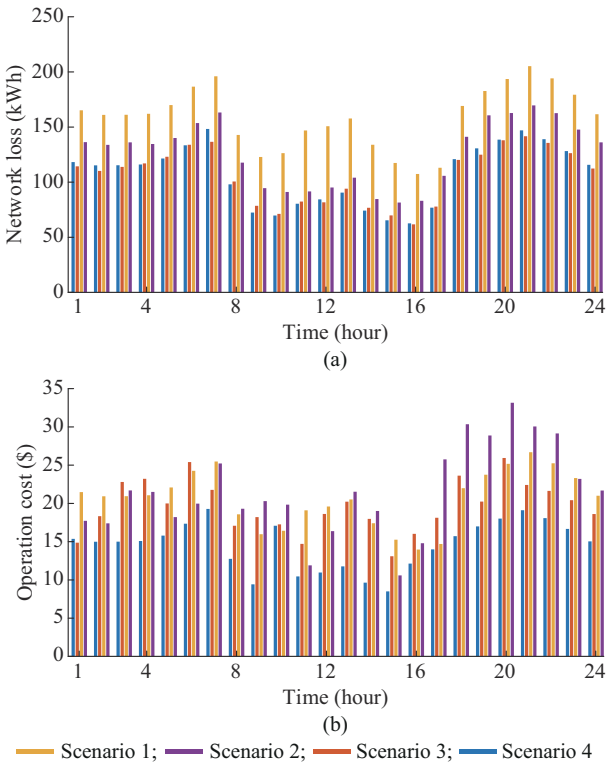


Fig. 15. Network loss and operation cost of different scenarios. (a) Network loss. (b) Operation cost.

The topology reverts to the disconnection state of branches 17-18 and 48-56. Therefore, we can conclude that the proposed method adjusts the control strategy according to the changes of system load and PV power, ensuring the economy and safety of system operation.

According to the above results, it can be concluded that the proposed method can reduce the operation cost of the complex three-phase unbalanced system with high PV penetration and avoid system overvoltage.

## V. Conclusion

This paper proposes a novel unbalanced DNR and SOP joint optimization control method, which translates SP into SP-MDP model based on MDP theory. Considering the large space of optimization decisions for three-phase unbalanced system, the BDDQN and MPSAC algorithms are developed based on the structural characteristics of DN. Furthermore, a DRL optimization method based on BDDQN and MPSAC combines the real-time system state to obtain the topology control strategy. Comprehensive test results on two unbalanced DNs show that the proposed BD-AC agent can effectively learn the reconfiguration and SOP joint control policy. Moreover, the data-driven SP method also reduces the operation cost of the DN, and relieves the problem of overvoltage, which has a much lower computation time than the model-based method.

With the extensive application of SOP in the future, SOP inevitably produces losses. Moreover, in practice, the line parameters of the DN are difficult to be determined accurately. Therefore, considering the SOP loss and the uncertainty of line parameters, it is our future research focus to propose a more accurate and robust DRL-based optimization method.

## Reference

[1] Y. Gao, W. Wang, J. Shi *et al*., "Batch-constrained reinforcement learning for dynamic distribution network reconfiguration," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5357-5369, Nov. 2020.

[2] M. Naguib, W. A. Omran, and H. E. A. Talaat, "Performance enhancement of distribution systems via distribution network reconfiguration and distributed generator allocation considering uncertain environment," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 3, pp. 647-655, May 2022.

[3] A. M. Eldurssi and R. M. O'Connell, "A fast nondominated sorting guided genetic algorithm for multi-objective power distribution system reconfiguration problem," *IEEE Transactions on Power Systems*, vol. 30, no. 2, pp. 593-601, Mar. 2015.

[4] F. Keynia, S. Esmaeili, and F. Sayadi, "Feeder reconfiguration and capacitor allocation in the presence of non-linear loads using new P-PSO algorithm," *IET Generation, Transmission & Distribution*, vol. 10, no. 10, pp. 2316-2326, Jul. 2016.

[5] X. Ji, Q. Liu, Y. Yu *et al.*, "Distribution network reconfiguration based on vector shift operation," *IET Generation, Transmission & Distribution*, vol. 12, no. 13, pp. 3339-3345, Jul. 2018.

[6] S. Zhang and M. Sridharan, "A survey of knowledge-based sequential decision making under uncertainty," *AI Magazine*, vol. 43, no. 2, pp. 1-6, Jun. 2022.

[7] Q. Zhang, Y. Kang, Y. Zhao *et al.*, "Traded control of human-machine systems for sequential decision-making based on reinforcement learning," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 4, pp. 553-566, Aug. 2022.

[8] L. Bai, T. Jiang, F. Li *et al.*, "Distributed energy storage planning in soft open point based active distribution networks incorporating network reconfiguration and DG reactive power capability," *Applied Energy*, vol. 210, pp. 1082-1091, Jan. 2018.

[9] R. You and X. Lu, "Voltage unbalance compensation in distribution feeders using soft open points," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 4, pp. 1000-1008, Jul. 2022.

[10] X. Dong, Z. Wu, G. Song *et al*., "A hybrid optimization algorithm for distribution network coordinated operation with SNOP based on simulated annealing and conic programming," in *Proceedings of 2016 IEEE PES General Meeting (PESGM)*, Boston, USA, Jul. 2016, pp. 1-5.

[11] M. B. Shafik, H. Chen, G. I. Rashed *et al.*, "Adequate topology for efficient energy resources utilization of active distribution networks equipped with soft open points," *IEEE Access*, vol. 7, pp. 99003-

99016, Jun. 2019.
[12] M. B. Shafik, G. I. Rashed, H. Chen *et al.*, "Reconfiguration strategy for active distribution networks with soft open points," in *Proceedings of 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, Xi'an, China, Jun. 2019, pp. 330-334.
[13] V. B. Pamshetti, S. Singh, and S. P. Singh, "Reduction of energy demand via conservation voltage reduction considering network reconfiguration and soft open point," *International Transactions on Electrical Energy Systems*, vol. 30, no. 1, pp. 1-8, Jan. 2020.
[14] I. Diaaeldin, S. Abdel Aleem, A. El-Rafei *et al.*, "Optimal network reconfiguration in active distribution networks with soft open points and distributed generation," *Energies*, vol. 12, no. 21, p. 4172, Nov. 2019.
[15] I. Sarantakos, N.-M. Zografou-Barredo, D. Huo *et al.,* "A reliability-based method to quantify the capacity value of soft open points in distribution networks," *IEEE Transactions on Power Systems*, vol. 36, no. 6, pp. 5032-5043, Nov. 2021.
[16] H. Ji, C. Wang, P. Li *et al.,* "An enhanced SOCP-based method for feeder load balancing using the multi-terminal soft open point in active distribution networks," *Applied Energy*, vol. 208, pp. 986-995, Dec. 2017.
[17] T. Ding, Z. Wang, W. Jia *et al.*, "Multiperiod distribution system restoration with routing repair crews, mobile electric vehicles, and soft-open-point networked microgrids," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 4795-4808, Nov. 2020.
[18] L. H. Macedo, J. F. Franco, M. J. Rider *et al.*, "Optimal operation of distribution networks considering energy storage devices," *IEEE Transactions on Smart Grid*, vol. 6, no. 6, pp. 2825-2836, Nov. 2015.
[19] J. Duan, D. Shi, R. Diao *et al.*, "Deep-reinforcement-learning-based autonomous voltage control for power grid operations," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 814-817, Jan. 2020.
[20] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE Journal of Power and Energy Systems*, vol. 4, no. 3, pp. 362-370, Sept. 2018.
[21] Z. Yan and Y. Xu, "Data-driven load frequency control for stochastic power systems: a deep reinforcement learning method with continuous action search," *IEEE Transactions on Power Systems*, vol. 34, no. 2, pp. 1653-1656, Mar. 2019.
[22] J. Jin and Y. Xu, "Optimal policy characterization enhanced actor-critic approach for electric vehicle charging scheduling in a power distribution network," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1416-1428, Mar. 2021.
[23] C. Wang, S. Lei, P. Ju *et al.*, "MDP-based distribution network reconfiguration with renewable distributed generation: approximate dynamic programming approach," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3620-3631, Jul. 2020.
[24] P. Li, H. Ji, C. Wang *et al.*, "Optimal operation of soft open points in active distribution networks under three-phase unbalanced conditions," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 380-391, Jan. 2019.
[25] X. Jiang, Y. Zhou, W. Ming *et al.*, "An overview of soft open points in electricity distribution networks," *IEEE Transactions on Smart Grid*, vol. 13, no. 3, pp. 1899-1910, May 2022.
[26] G. Carpinelli, G. Celli, S. Mocci *et al.*, "Optimal integration of distributed energy storage devices in smart grids," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 985-995, Jun. 2013.
[27] T. Yang, Y. Guo, L. Deng *et al.*, "A linear branch flow model for radial distribution networks and its application to reactive power optimization and network reconfiguration," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2027-2036, May 2021.
[28] M. Van Otterlo and M. Wiering, "Reinforcement learning and Markov decision processes," in *Reinforcement Learning*. New York: Springer, 2012, pp. 3-42.
[29] S. Wang, L. Du, X. Fan *et al.*, "Deep reinforcement scheduling of energy storage systems for real-time voltage regulation in unbalanced LV networks with high PV penetration," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 4, pp. 2342-2352, Oct. 2021.
[30] Y. Du and F. Li, "Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1066-1076, Mar. 2020.
[31] S. Wang, R. Diao, C. Xu *et al.*, "On multi-event co-calibration of dynamic model parameters using soft actor-critic," *IEEE Transactions on Power Systems*, vol. 36, no. 1, pp. 521-524, Jan. 2021.
[32] Q. Huang, X. Xu, F. Blaabjerg *et al.*, "Deep reinforcement learning based approach for optimal power flow of distribution networks embedded with renewable energy and storage devices," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 5, pp. 1101-1110, Sept. 2021.
[33] A. Tavakoli, F. Pardo, and P. Kormushev, "Action branching architectures for deep reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, USA, Jul. 2018, pp. 4131-4138.
[34] IEEE. (2021, Jun.). Resources | PES test feeder. [Online]. Available: https://site.ieee.org/pes-testfeeders/resources/
[35] X. Ji, Z. Yin, Y. Zhang *et al.*, "Real-time robust forecasting-aided state estimation of power system based on data-driven models," *International Journal of Electrical Power & Energy Systems*, vol. 125, pp. 1-11, Feb. 2021.
[36] T. Hong, P. Pinson, S. Fan *et al.*, "Probabilistic energy forecasting: global energy forecasting competition 2014 and beyond," *International Journal of Forecasting*, vol. 32, no. 3, pp. 896-913, Jul. 2016.
[37] Y. Zhang, X. Ji, J. Xu *et al.*, "Dynamic reconfiguration of distribution network based on temporal constrained hierarchical clustering and fireworks algorithm," in *Proceedings of 2020 IEEE/IAS Industrial and Commercial Power System Asia*, Weihai, China, Jul. 2020, pp. 1702-1708.
[38] H. Zhai, M. Yang, B. Chen *et al.*, "Dynamic reconfiguration of three-phase unbalanced distribution networks," *International Journal of Electrical Power & Energy Systems*, vol. 99, pp. 1-10, Jul. 2018.
[39] J. Xiao, Y. Li, X. Qiao *et al.*, "Enhancing hosting capacity of uncertain and correlated wind power in distribution network with ANM strategies," *IEEE Access*, vol. 8, pp. 189115-189128, Jan. 2020.

**Ziyang Yin** received the B.S. and M.S. degrees in electrical engineering from Shandong University of Science and Technology, Qingdao, China, in 2018 and 2021, respectively. He is currently pursuing the Ph.D. degree in School of Electrical and Information Engineering, Tianjin University, Tianjin, China. His research interests include distributed generation, machine learning, and smart distribution system.

**Shouxiang Wang** received the B.S. and M.S. degrees in electrical engineering from Shandong University of Technology, Jinan, China, in 1995 and 1998, respectively, and the Ph.D. degree in electrical engineering from Tianjin University, Tianjin, China, in 2001. He is currently a Professor with the School of Electrical and Information Engineering, and Deputy Director of Key Laboratory of Smart Grid of Ministry of Education, Tianjin University. His research interests include distributed generation, microgrid, and smart distribution system.

**Qianyu Zhao** received the B.S. and M.S. degrees in electrical engineering and control science and engineering from Tiangong University, Tianjin, China, in 2012 and 2015, respectively, and the Ph.D. degree in electrical engineering from Tianjin University, Tianjin, China, in 2020. She is currently an Assistant Professor with the School of Electrical and Information Engineering, Tianjin University. Her research interests include planning, assessment of energy storage and distributed generation, uncertainty analysis of distribution networks.