

Reinforcement Learning for Active Distribution Network Planning Based on Monte Carlo Tree Search

Xi Zhang^a, Weiqi Hua^b, Youbo Liu^{a*}, Jiajun Duan^c, Zhiyuan Tang^a, Junyong Liu^a

^aCollege of Electrical Engineering, Sichuan University, Sichuan 610065, China

^bOxford e-Research Centre, Department of Engineering Science, University of Oxford OX1 3QG, UK

^cNextracker Inc., Fremont, CA94555, USA

Abstract

Active distribution network planning is of importance for utility companies in terms of distributed generation investment, reliability assessment, optimal reactive power planning, substation evaluation, and feeder reconfiguration. However, it is challenging for current model-based optimization problems to guarantee the performances of active distribution network planning, due to an empirically pre-defined solution space. To overcome this issue, this paper proposes a performance-oriented method for the active distribution network planning. The solution space of the planning model is dynamically updated through using deep neural networks which are trained by the Monte Carlo tree search-based reinforcement learning until the desired performances are satisfied. **Simulation results based on the standard IEEE 33-bus test system demonstrate that the proposed method can successfully improve the performances of the active distribution network planning to a desired level at a lower investment cost compared to other cases.**

© 2017 Elsevier Inc. All rights reserved.

Keywords: Active Distribution Network Planning; Convex Optimization; Monte Carlo tree search; Reinforcement Learning; Renewable Energy Source

1. Introduction

Active distribution networks (ADNs) planning refers to that a utility company upgrades the traditional feeders and substations and strategically makes the investment decisions on its portfolio of generation technologies (e.g., the energy storage system (ESS) [1-3], reactive power sources [4, 5], and distributed generations [6-8]), locations, and build rate, to guarantee certain performances, e.g., reducing carbon emissions [9, 10] and minimizing the amount of the unserved load (AUL) [11, 12]. With the aim of net zero carbon emissions and clean energy supply, facilitating the penetration of renewable energy sources (RESs) is an important performance that needs to be achieved by the ADN planning [1, 4, 15]. Research in [1] and [15] used the ESS to maximize the size of the renewable power absorbed by the ADN. In [4], the SVC devices were adopted to enhance the hosting capacity of the photovoltaic (PV) generation. Enhancing the system reliability is another important issue which was discussed by [16-18], where the ESS [16], distribution automation systems [17], and distribution switches [18] were optimally configured to improve the system reliability. In [19], the optimal mix, siting, and sizing of the wind turbine, PV, and ESS were identified to maximize the profits of the ADN. Research in [20] optimizes the settings of the ADN management technologies to minimize the total investment and operational costs, in which the RES curtailment and load curtailment were considered in the operational cost. Research in [21] and [22] specifically focused on improving the resilience of the distribution networks. Reference [23] optimally reinforced the ADN facilities to satisfy the increasing demand of the electric vehicle charging. Reference [24] adopted the RES, ESS, and demand response to alleviate carbon emissions. However, current optimization-based approaches empirically predefine a combination of the ADN management technologies as a fixed space, so as to explore optimal solutions within the space. These performances could be potentially further improved when the solution spaces are dynamically updated.

Reinforcement learning (RL) have drawn attentions in solving the ADN problems with dynamic solution spaces, since it is a model-free approach and is capable of adapting with environments to make optimal decisions. Implementing the RL to solve the problems of the ADN management has been well documented in literature. Research

in [25] proposed a multi-agent deep reinforcement learning (DRL) approach for the voltage regulation, through using the on-load tap changers, CBs, and PV inverters. Researchers in [26] proposed an on-line and off-line volt/var control strategy to improve the voltage profile by regulating the inverter-based energy resources through using the DRL. Research in [27] provided a two-timescale voltage control which cooperatively adjusts the CBs and smart inverters. In [28], a constrained soft actor-critic algorithm was implemented to achieve a safe volt/var control. Research in [29] proposed a batch-constrained reinforcement learning for dynamic distribution network reconfiguration using a finite historical operational dataset. Reference in [30] implemented the RL to schedule the operation of an energy hub consisting of electrical and natural gas networks. With respect to solving sequential decision-making problems, the Monte Carlo tree search-based reinforcement learning (MCTS-RL) [31, 32] becomes a promising solution, given its ability to find potential optimal solutions through adapting state variations. Research in [31] adopted the MCTS-RL to mitigate the overvoltage issue in distribution systems with the high PV penetration. In [32], the MCTS-RL was used to schedule the stochastic maintenance of the ADN. There is still great opportunity for dynamically exploring the solution space (SP) of the ADN planning problems through using the MCTS-RL.

This paper proposes a novel method for ADN planning through using the MCTS-RL to yield an ADN planning scheme (APS), under which the desired performances are guaranteed through dynamically updating the solution spaces. The main contributions of this paper are summarized as follows:

- Instead of solving the model-based optimization problem with fixed solution spaces, this paper models the dynamic update of the solution spaces of the APS as a Markov decision-making process (MDP) to guarantee the desired performances.
- The MCTS-RL is implemented to train the deep neural networks (DNNs) to guarantee that the APS meets the desired performances.
- The proposed method is tested on the standard IEEE 33-bus system and the numerical results demonstrate that the expected performance of the ADN can be guaranteed at a lower investment cost compared to other cases.

The rest of this paper is organized as follows. Section 2 describes the MDP process for ADN planning. Section 3 introduces the methodology for obtaining the optimal planning scheme. Section 4 provides case studies to validate the effectiveness of the proposed model. Section 5 draws conclusions and lists the future work.

Nomenclature

ADN	Active distribution network.
APS	ADN planning scheme.
ESS	Energy storage system.
SVC	Static var compensator.
CB	Capacitor bank.
RES	Renewable energy sources.
c^{loss}	Cost of network loss.
c^{res}	Cost of cutting renewable energy sources.
c^{load}	Cost of cutting load.
r_{ij}, x_{ij}	Resistance and reactance of a branch between bus i and bus j .
η	Ratio of reactive power to active power.
$P_j^{\text{tr,max}} \setminus Q_j^{\text{tr,max}}$	Maximum active/reactive power injections by transformers.
$P_j^{\text{ess,max}}$	Maximum active power generated by ESS.
$E_j^{\text{ess,min}} \setminus E_j^{\text{ess,max}}$	Minimum/maximum state of charge of ESS.
$P_j^{\text{gas,min}} \setminus P_j^{\text{gas,max}}$	Minimum/maximum active power of gas generator.
$Q_j^{\text{svc,min}} \setminus Q_j^{\text{svc,max}}$	Minimum/maximum reactive power of SVC.
$Q_j^{\text{cb,min}} \setminus Q_j^{\text{cb,max}}$	Maximum active power generated by ESS.
\mathcal{W}	Set of scenarios.
\mathcal{T}	Set of time intervals.
\mathcal{E}	Set of branches in distribution network.
\mathcal{N}^{B}	Set of nodes in distribution network.

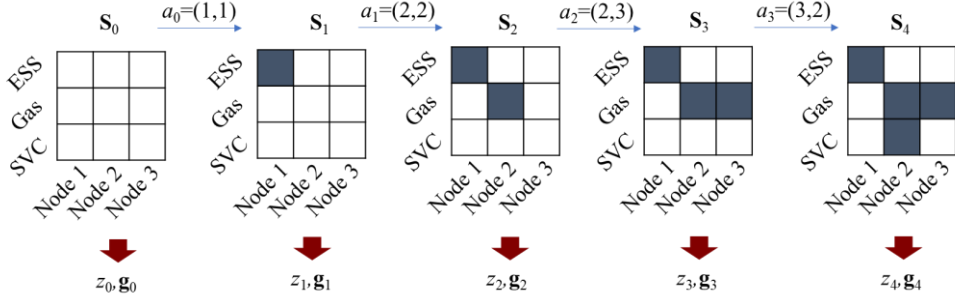


Fig. 2. MDP process of the distribution network planning.

2.2. Optimization Problem for Active Distribution Network Planning

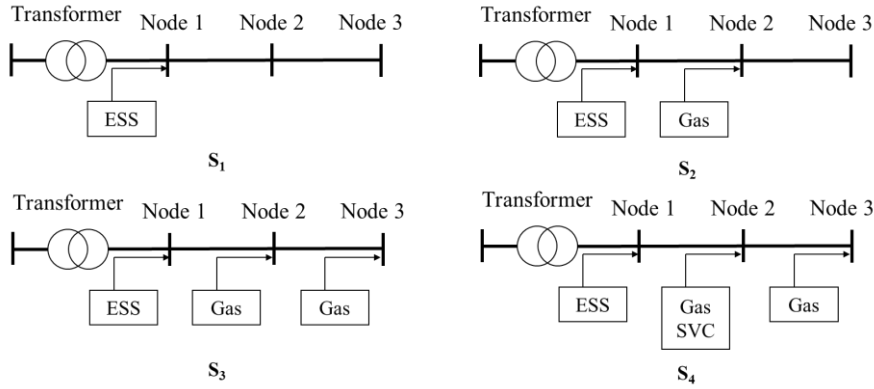


Fig. 3. State transition of 3-node system representing the active distribution network planning scheme.

The state transition of a 3-node system is presented in Fig. 3. Different planning schemes would yield different performances. In this research, the ADN performances are instantiated by the hosting capacity of RES and amount of unserved loads. To evaluate these performances at every state, each state is modelled as a convex optimization problem as follows:

$$\min : f(s_j^h, \Delta P_{i,t,w}^{\text{res}}, \Delta P_{i,t,w}^{\text{aul}}) = \sum_{j \in \mathcal{N}^b} \sum_{h \in \mathcal{H}} c_j^{h, \text{inv}} s_j^h + \sum_{w \in \mathcal{W}} \sum_{t \in \mathcal{T}} \sum_{ij \in \mathcal{E}} c^{\text{loss}} \tilde{I}_{ij,t,w} r_{ij} + \sum_{w \in \mathcal{W}} \sum_{t \in \mathcal{T}} \sum_{ij \in \mathcal{E}} c_i^{\text{res}} \Delta P_{i,t,w}^{\text{res}} + \sum_{w \in \mathcal{W}} \sum_{t \in \mathcal{T}} \sum_{ij \in \mathcal{E}} c_i^{\text{aul}} \Delta P_{i,t,w}^{\text{aul}} \quad (1)$$

s.t.

$$\sum_{k \in \delta(j)} P_{jk,t,w} - \sum_{i \in \delta(j)} (P_{ij,t,w} - \tilde{I}_{ij,t,w} r_{ij}) = P_{j,t,w}^{\text{tr}} + P_{j,t,w}^{\text{ess}} + P_{j,t,w}^{\text{gas}} + (P_{j,t,w}^{\text{res}} - \Delta P_{i,t,w}^{\text{res}}) + (P_{j,t,w}^{\text{load}} - \Delta P_{i,t,w}^{\text{aul}}) \quad (2)$$

$$\sum_{k \in \delta(j)} Q_{jk,t,w} - \sum_{i \in \delta(j)} (Q_{ij,t,w} - \tilde{I}_{ij,t,w} x_{ij}) = Q_{j,t,w}^{\text{tr}} + Q_{j,t,w}^{\text{sbc}} + Q_{j,t,w}^{\text{cb}} + (Q_{j,t,w}^{\text{load}} - \eta \Delta P_{i,t,w}^{\text{res}}) \quad (3)$$

$$\tilde{V}_{j,t,w} = \tilde{V}_{i,t,w} - 2(P_{ij,t,w} r_{ij} + Q_{ij,t,w} x_{ij}) + \tilde{I}_{ij,t,w} (r_{ij}^2 + x_{ij}^2) \quad (4)$$

$$\tilde{V}_{i,t,w} = V_{i,t,w}^2 \quad (5)$$

$$\tilde{I}_{ij,t,w} = \frac{P_{ij,t,w}^2 + Q_{ij,t,w}^2}{V_{i,t,w}^2}, \quad (6)$$

$$\left\| \begin{array}{c} 2P_{ij,t,w} \\ 2Q_{ij,t,w} \\ \tilde{I}_{ij,t,w} - \tilde{V}_{i,t,w} \end{array} \right\| \leq \tilde{I}_{ij,t,w} + \tilde{V}_{i,t,w}, \quad (7)$$

$$0 \leq \tilde{I}_{ij,t,w} \leq (I_{ij,t,w}^{\max})^2, \quad (8)$$

$$(V_{i,t,w}^{\min})^2 \leq \tilde{V}_{i,t,w} \leq (V_{i,t,w}^{\max})^2, \quad (9)$$

$$\Delta P_{i,t,w}^{\text{res}} \leq P_{j,t,w}^{\text{res}}, \quad (10)$$

$$\Delta P_{i,t,w}^{\text{paul}} \leq P_{j,t,w}^{\text{load}}, \quad (11)$$

$$\begin{cases} P_{j,t,w}^{\text{tr,min}} \leq P_{j,t,w}^{\text{tr}} \leq P_{j,t,w}^{\text{tr,max}} \\ Q_{j,t,w}^{\text{tr,min}} \leq Q_{j,t,w}^{\text{tr}} \leq Q_{j,t,w}^{\text{tr,max}} \end{cases}, \quad (12)$$

$$\begin{cases} P_{j,t,w}^{\text{ess}} = P_{j,t,w}^{\text{discharge}} - P_{j,t,w}^{\text{charge}} \\ 0 \leq P_{j,t,w}^{\text{discharge}} \leq s_j^{\text{ess}} u_{j,t,w}^{\text{discharge}} P_j^{\text{ess,max}} \\ 0 \leq P_{j,t,w}^{\text{charge}} \leq s_j^{\text{ess}} u_{j,t,w}^{\text{charge}} P_j^{\text{ess,max}} \\ E_{j,t+1,w}^{\text{ess}} = E_{j,t,w}^{\text{ess}} + \alpha_j^{\text{charge}} P_{j,t,w}^{\text{charge}} - \alpha_j^{\text{discharge}} P_{j,t,w}^{\text{discharge}}, \quad j \in \mathcal{N}^{\text{E}}, \\ E_j^{\text{ess,min}} \leq E_{j,t,w}^{\text{ess}} \leq E_j^{\text{ess,max}} \\ u_{j,t,w}^{\text{charge}} + u_{j,t,w}^{\text{discharge}} \leq 1 \end{cases}, \quad (13)$$

$$s_j^{\text{gas}} P_{j,t,w}^{\text{gas,min}} \leq P_{j,t,w}^{\text{gas}} \leq s_j^{\text{gas}} P_{j,t,w}^{\text{gas,max}}, \quad j \in \mathcal{N}^{\text{G}}, \quad (14)$$

$$s_j^{\text{svc}} Q_{j,t,w}^{\text{svc,min}} \leq Q_{j,t,w}^{\text{svc}} \leq s_j^{\text{svc}} Q_{j,t,w}^{\text{svc,max}}, \quad j \in \mathcal{N}^{\text{S}}, \quad (15)$$

$$s_j^{\text{cb}} Q_{j,t,w}^{\text{cb,min}} \leq Q_{j,t,w}^{\text{cb}} \leq s_j^{\text{cb}} Q_{j,t,w}^{\text{cb,max}}, \quad j \in \mathcal{N}^{\text{C}}. \quad (16)$$

Equation (1) describes the objective function of annual operational costs, in which the first term is the cost of the network loss, the second term is the cost of cutting RES power, and the third term represents the cost of the load curtailment. In this paper, we have $\mathcal{H} = \{\text{ESS, Gas, SVC, CB}\}$. Equations (2)-(7) are the convexified AC power flow constraints, where equations (2) and (3) are the network power flow constraints, equations (4) and (5) denote the voltage constraints, equation (6) is the branch current constraint, equation (7) is the relaxed second-order cone constraint. Equations (8) and (9) are the security constraints that are used to confine the square of the current and voltage, respectively. Equations (10) and (11) are the constraints for cutting renewables and loads. Equation (12) is the active and reactive power constraints for the transformer, and equation (13) is the constraint for the charging and discharging behaviors of the ESS. Equations (14)-(16) are the output range of the gas generators, SVC, and CB, respectively. s_j^h is a binary variable which is used to indicate whether the ADN management technology h is installed on the node j ($s_j^h=1$) or not ($s_j^h=0$). \mathcal{N}^{E} , \mathcal{N}^{G} , \mathcal{N}^{S} , and \mathcal{N}^{C} are the sets of the candidate nodes for installing ESS, gas generator, SVC and CB.

A new state \mathbf{S}_n would result in an updated solution space as defined by equations (13)-(16) and corresponding optimization problem consisting of equations (1)-(16). Let A_n denote this optimization model. Through solving the optimization problem A_n , the value used to evaluate the quality of an ASP (z_n) and performance index vector (\mathbf{g}_n) can be obtained as

$$z_n = \begin{cases} \frac{H}{\sum_{j \in \mathcal{N}^{\text{B}}} \sum_h c^{h,\text{inv}} s_j^h}, & h \in \{\text{ESS, Gas, SVC, CB}\}, \quad (A_n \text{ is feasible}) \\ -M, & (A_n \text{ is infeasible}) \end{cases}, \quad (17)$$

$$\mathbf{g}_n \begin{cases} [g_1, g_2] = [\sum_{w \in W} \sum_{t \in T} \sum_{i \in N^B} \Delta P_{i,t,w}^{\text{res}}, \sum_{w \in W} \sum_{t \in T} \sum_{i \in N^B} \Delta P_{i,t,w}^{\text{aul}}], & (A_n \text{ is feasible}) \\ [0, 0], & (A_n \text{ is infeasible}) \end{cases}, \quad (18)$$

where the denominator in equation (17) indicates the total investment cost, $c^{h,\text{inv}}$ is the investment cost of installing the device h , H is a constant determining the range of z_n , and **M is a very large number**. In equation (18), g_1 is the total RES power curtailment and g_2 is the AUL.

3. Monte Carlo Tree Search-Based Reinforcement Learning

This section introduces the developed MCTS-RL algorithm for exploring the desired performances of the ADN.

3.1. Key Concepts

To facilitate the illustration of our proposed method, three key concepts related to the MCTS-RL are introduced as follows:

- *Root node*: the starting node of the Monte Carlo tree, denoted as \mathbf{S}_0 .
- *Leaf node*: the last node that can be currently visited in the Monte Carlo tree, denoted as \mathbf{S}_l .
- *Policy network*: a DNN that is used to predict the value of a state \mathbf{S} and the probabilities for all the legal actions belonging to the state \mathbf{S} .

3.2. Monte Carlo Tree Search Algorithm

The process of the MCTS is presented in Fig. 3. The basic component of the MCTS is the node which represents a state \mathbf{S} . The nodes store edges for all the legal actions. Each edge (\mathbf{S}, \mathbf{a}) represents the execution of an action \mathbf{a} at the current state \mathbf{S} . The edge contains four statistics, i.e., $N(\mathbf{S}, \mathbf{a})$, $W(\mathbf{S}, \mathbf{a})$, $Q(\mathbf{S}, \mathbf{a})$, and $P(\mathbf{S}, \mathbf{a})$, where $N(\mathbf{S}, \mathbf{a})$ is the count of selecting the action \mathbf{a} under the state \mathbf{S} , $W(\mathbf{S}, \mathbf{a})$ is the total value of selecting the action \mathbf{a} under the state \mathbf{S} , $Q(\mathbf{S}, \mathbf{a})$ is the mean value of selecting the action \mathbf{a} under the state \mathbf{S} , and $P(\mathbf{S}, \mathbf{a})$ is the prior probability of selecting the action \mathbf{a} under the state \mathbf{S} . We have $Q(\mathbf{S}, \mathbf{a}) = W(\mathbf{S}, \mathbf{a}) / P(\mathbf{S}, \mathbf{a})$. The main process of the MCTS contains the following three steps:

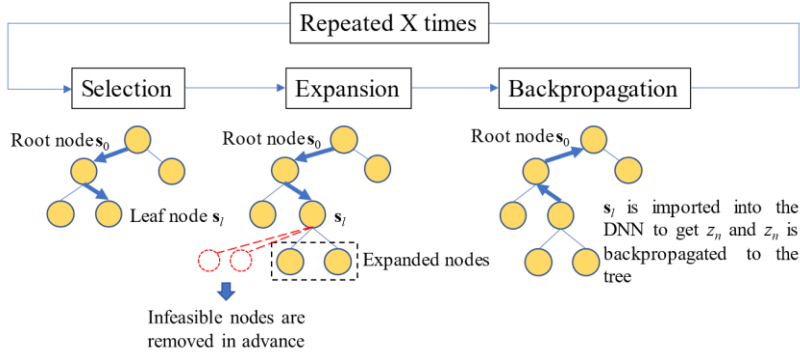


Fig. 4. Process of the Monte Carlo tree search algorithm. The selection step is used to select the next node. The expansion step is used to formulate all the next node. The backpropagation step is used to update the parameters that belong to each node.

Step 1) Selection.

The traversal of the MCTS starts from the root node \mathbf{S}_0 and iteratively selects the next node according to equation (19) and ends at a leaf node \mathbf{S}_l which has not been expanded yet.

$$\mathbf{a}_n = \arg \max_{a \in \mathcal{L}_{s_n}} (Q(\mathbf{S}_n, \mathbf{a}) + U(\mathbf{S}_n, \mathbf{a})), \quad n \in \mathcal{R}_{s_n \rightarrow s_l} \quad (19)$$

where \mathbf{a}_n is the action selected under the current state \mathbf{S}_n . $\mathcal{R}_{s_n \rightarrow s_l}$ is the set of all the nodes in the path that starts from

\mathbf{S}_n and ends at \mathbf{S}_l . $\mathcal{L}_{\mathbf{S}_n}$ is the set of all the legal actions at the state \mathbf{S}_n . $U(\mathbf{S}_n, \mathbf{a})$ is the upper confidence bound which is calculated as [35].

$$U(\mathbf{S}_n, \mathbf{a}) = c_{\text{puct}} P(\mathbf{S}_n, \mathbf{a}) \frac{\sqrt{\sum_b N(\mathbf{S}_p, \mathbf{b})}}{1 + N(\mathbf{S}_n, \mathbf{a})}, \quad (20)$$

where c_{puct} is a constant determining the level of exploration, \mathbf{S}_p is the parent node of \mathbf{S}_n , $\sum_b N(\mathbf{S}_p, \mathbf{b})$ is the total visits of the parent node \mathbf{S}_p , \mathbf{b} is the available action of \mathbf{S}_p , and $P(\mathbf{S}_n, \mathbf{a})$ is the output of the policy network.

Step 2) Expansion

When a leaf node \mathbf{S}_l is encountered, the state \mathbf{S}_l is imported into the policy network. The probabilities of the actions and the value of the state \mathbf{S}_l are obtained as

$$(\mathbf{p}_l, v_l, \mathbf{k}_l) = f_\theta(\mathbf{S}_l), \quad (21)$$

where \mathbf{p}_l is the vector of the probabilities for all the legal actions belonging to the node \mathbf{S}_l , v_l is the value of the node \mathbf{S}_l , and \mathbf{k}_l is the performance index set generated by the policy network which is used to judge the terminal state, f_θ is the policy network. **It is noted that the value v_l is yielded by training the policy network, whereas the value z_n is yielded by solving the optimization problem. The difference between v_l and z_n indicates how far the training output is deviated from the theoretical one.** Then, the leaf node \mathbf{S}_l is expanded and each edge $(\mathbf{S}_l, \mathbf{a})$ is initialized as

$$\begin{cases} N(\mathbf{S}_l, \mathbf{a}) = 0, W(\mathbf{S}_l, \mathbf{a}) = 0 \\ Q(\mathbf{S}_l, \mathbf{a}) = 0, P(\mathbf{S}_l, \mathbf{a}) = p_a, p_a \in \mathbf{p} \end{cases}. \quad (22)$$

Step 3) Backpropagation

After the leaf node \mathbf{S}_l is expanded, the edge statistics of the node \mathbf{S}_n in the searching trajectory $\mathcal{R}_{\mathbf{S}_n \rightarrow \mathbf{S}_l}$ are updated by

$$\begin{cases} N(\mathbf{S}_n, \mathbf{a}_n)^{\text{new}} = N(\mathbf{S}_n, \mathbf{a}_n)^{\text{old}} + 1 \\ W(\mathbf{S}_n, \mathbf{a}_n)^{\text{new}} = W(\mathbf{S}_n, \mathbf{a}_n)^{\text{old}} + v_l, \\ Q(\mathbf{S}_n, \mathbf{a}_n)^{\text{new}} = \frac{W(\mathbf{S}_n, \mathbf{a}_n)^{\text{new}}}{N(\mathbf{S}_n, \mathbf{a}_n)^{\text{new}}} \end{cases} \quad (23)$$

The algorithm of the MCTS is shown in **Algorithm I**.

Algorithm I: MCTS

- 1) **Input:** the current state \mathbf{S}_n
 - 2) **for** counter 1 to n^{search} , **do**:
 - 3) **while** \mathbf{S}_n is not a leaf node, **do**:
 - 4) Select the next action \mathbf{a}_n according to equations (19) and (20).
 - 5) Execute \mathbf{a}_n and get the next state \mathbf{S}^{next}
 - 6) $\mathbf{S}_n \leftarrow \mathbf{S}^{\text{next}}$
 - 7) Obtain \mathbf{p}_l , v_l , and \mathbf{k}_l according to equation (21).
 - 8) **if** \mathbf{S}_l is not a terminal state (according to \mathbf{k}_l) **do**:
 - 9) Expand \mathbf{S}_l
 - 10) Backup from \mathbf{S}_l using v_l
 - 11) **end**
 - 12) **Output:** the search probabilities $\pi(\mathbf{a}|\mathbf{S}_n)$.
-

3.3. Self-Play Algorithm

The self-play algorithm is capable of training the policy network to accurately predict optimal decisions, which has been demonstrated by the AlphaZero [34]. The self-play starts from a root state \mathbf{S}_0 and ends at a terminal state \mathbf{S}_e . When executing the next move, several rounds of the MCTS searching (*selection* \rightarrow *expansion* \rightarrow *backup*) from current state \mathbf{S}_n ($\mathbf{S}_n \in \mathcal{R}_{\mathbf{S}_n \rightarrow \mathbf{S}_l}$) are firstly conducted. Then an actual probability distribution $\pi(\mathbf{a}|\mathbf{S}_n)$ over the legal actions of the \mathbf{S}_n is calculated by

$$\pi(\mathbf{a} | \mathbf{S}_n) = \frac{N(\mathbf{S}_n, \mathbf{a})^{1/\tau}}{\sum_b N(\mathbf{S}_n, \mathbf{b})^{1/\tau}}, \quad (24)$$

where τ is a temperature factor that is used to weigh the exploration and exploitation. The next move is selected by

$$\mathbf{a}_n^{\text{play}} = \arg \max_{\mathbf{a} \in L(\mathbf{s}_n)} \pi(\mathbf{a} | \mathbf{S}_n). \quad (25)$$

The difference between equation (19) and equation (25) is that equation (19) selects the next action based on the probabilities produced by the policy network while equation (25) selects the next action according to the probability of the next move which is proportional to its exponentiated visit count. When a terminal state \mathbf{S}_e is encountered, one round of the game stops and a value z_n is obtained based on the game rules and returns

$$\{(\mathbf{S}_n, \pi(\mathbf{a} | \mathbf{S}_n), z_n, \mathbf{g}_n) | n \in \mathcal{R}_{\mathbf{S}_0 \rightarrow \mathbf{S}_e}\}, \quad (26)$$

where \mathbf{g}_n and z_n are obtained by solving A_n generated by \mathbf{S}_n . The self-play algorithm is shown in **Algorithm II**.

Algorithm II: self-play

- 1) **Input:** the root state \mathbf{S}_0
 - 2) $\mathbf{S}_n \leftarrow \mathbf{S}_0$
 - 3) **while:**
 - 4) Evaluate \mathbf{S}_n by optimizing the model of equations (1)-(16)
 - 5) Get \mathbf{g}_n and z_n according to equations (17) and (18)
 - 6) **if** $\mathbf{g}_n \leq \mathbf{g}^{\text{expected}}$ (\mathbf{g}_n satisfies the expected performance):
 - 7) Return the episode data as shown in equation (26).
 - 8) Start MCTS searching from \mathbf{s}_n (**Algorithm I**) and get $\pi(\mathbf{a} | \mathbf{S}_n)$
 - 9) Select the next action \mathbf{a}_n according to (24) and (25).
 - 10) Execute \mathbf{a}_n and get the next state \mathbf{S}^{next}
 - 11) $\mathbf{S}_n \leftarrow \mathbf{S}^{\text{next}}$
 - 12) **End**
 - 13) **Output:** one episode data as presented in (26).
-

3.4. Training Policy Network

The parameters of the policy network θ are updated by minimizing the loss function as

$$l^{\text{new}} = \frac{1}{D} \sum_{d \in D} (z_d - v_d)^2 - \sum_{d \in D} \pi(\mathbf{a} | \mathbf{S}_d)^T \log \mathbf{p}_d + \frac{1}{D} \sum_{d \in D} (\mathbf{g}_d - \mathbf{k}_d)^T (\mathbf{g}_d - \mathbf{k}_d) + c \|\theta\|^2, \quad (27)$$

where D is the size of one batch of data sampled from the memory \mathcal{M} , z_d is the value of the state \mathbf{S}_d obtained through solving the optimization problem, v_d is the predicted value generated by the policy network, π_d is the probability distribution over the legal actions under \mathbf{S}_d , \mathbf{p}_d is the predicted action probability under \mathbf{S}_d . Note that z_d and π_d are obtained by the self-play algorithm. v_d and \mathbf{p}_d are the outputs of the policy network used to guide the MCTS search. The algorithm for training the policy network is presented in **Algorithm III**.

Algorithm III: training policy network

- 1) **Input:** \mathbf{S}_l
 - 2) Initialize the parameters of the policy network to randomly weights θ_0 .
 - 3) **for** counter 1 to the number of training episodes n^{eps} **do:**
 - 4) Start self-play (**Algorithm II**) and get one episode data.
 - 5) Append the data to the memory \mathcal{M}
 - 6) **if** the length of memory is greater than the batch size D , **do:**
 - 7) Sample one batch of data with size D from the memory \mathcal{M} .
 - 8) Update the parameters of the policy network by minimizing (27)
 - 9) **End**
 - 10) **Output:** \mathbf{p}_l , v_l , and \mathbf{k}_l
-

4. Numerical Results

In this section, case studies have been conducted to validate the proposed model under various cases of technological combinations.

4.1. Simulation Setup

The proposed algorithms were written in Python 3.6.5 using Keras 2.3.1 backend by Tensorflow 1.14.0 and executed on a machine with the Intel Core i6 1.6GHz with 8GB RAM. GUROBI 9.0 is used to solve the optimization model. The standard IEEE 33-bus distribution network (see Fig. 5) is used to verify the proposed method. The sequential load demand and PV generation in four typical scenarios during a year is presented in Fig. 4. The nodes 9, 11, 14, 17, 19, 21, 23, and 29 in the IEEE 33-bus system are equipped with 0.5MW PV.

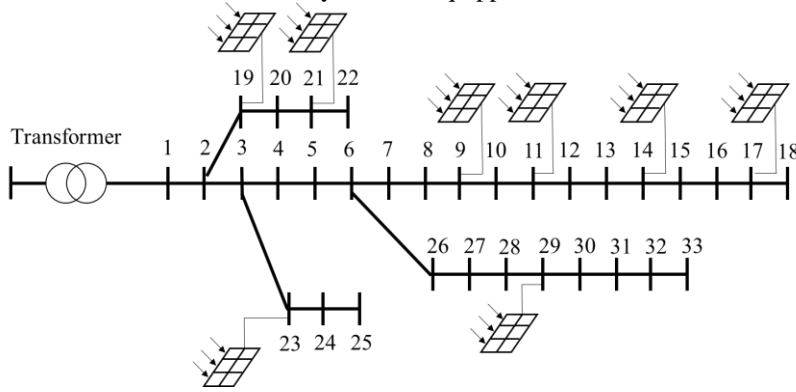


Fig. 5. Schematic illustration of the IEEE 33-bus system. The nodes 9, 11, 14, 17, 19, 21, 23, and 29 in the IEEE 33-bus system are equipped with 0.5MW PV.

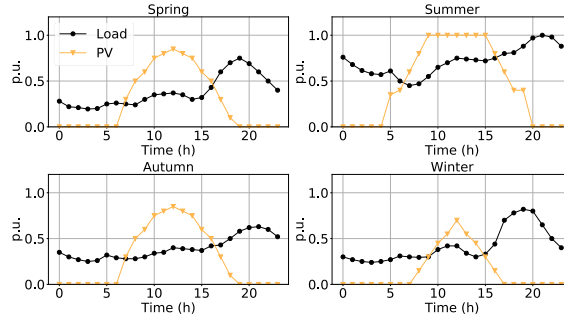


Fig. 6. Four typical scenarios of daily load and PV profiles.

The parameters in [11] and [34] are used in our research with details listed in Table 1.

Table 1. Parameters used in this research.

Type	Name	Value
MCTS	C_{puct}	5.0
	τ	1.0
	n^{search}	500
Policy network	c	10^4

	n^{eps}	400
	Input layer dimension	33×4
	The first hidden layer dimension	128
	The second hidden layer dimension	128
Output layer dimension	Channel 1	33×4
	Channel 2	1
	Channel 3	2×1
	Learning rate	0.0001
	Memory size	10000
	Batch size	32
Optimization model	c^{Loss}	200\$/MWh
	w	3000
	c^{RES}	500\$/MW
	c^{AUL}	5000\$/MW
	$c^{\text{SVC,INV}}$	0.5e6\$/MVar
	$c^{\text{CB,INV}}$	40000\$/MVar
	$c^{\text{Gas,INV}}$	30000\$/MW
	$c^{\text{ESS,INV}}$	200000\$/MW
	$p^{\text{ESS,max}}$	0.1MW
	$E^{\text{ESS,max}}$	0.5MWh
	$E^{\text{ESS,min}}$	0.1MWh
	α^{charge}	0.9
	$\alpha^{\text{discharge}}$	1.1
	$p^{\text{Gas,max}}$	0.5MW
	$Q^{\text{SVC,max}}$	0.5MVar
	$Q^{\text{CB,max}}$	0.5Mvar
	g^{expected}	[0.1, 0.1]

4.2. System Performance

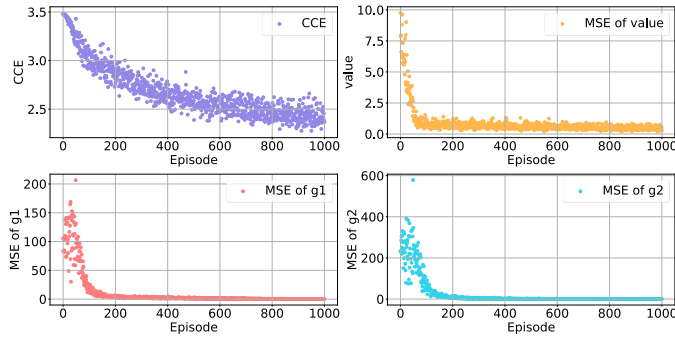


Fig. 6. The performances and losses with the increase of the training iterations.

The losses for three terms in equation (27) are presented in Fig. 6. It can be seen that the losses decrease as the training continues, which means that the policy network can accurately predict the action probability, the value, the amount of the load and PV curtailment within the acceptable level (1×10^{-6}). The well-trained policy network is then used to generate the APS from the initial state \mathbf{S}_0 (the elements in \mathbf{S}_0 are zero).

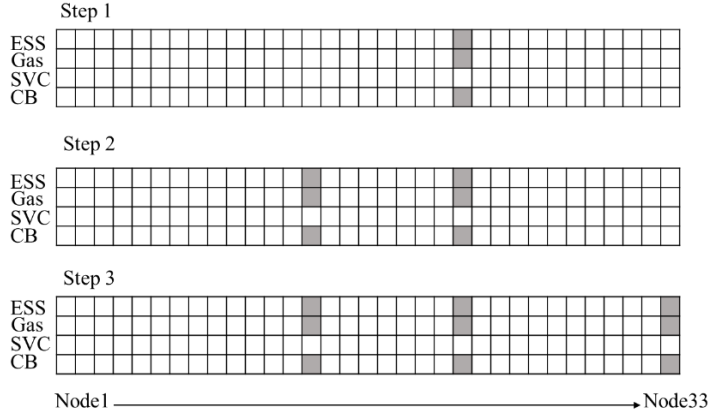


Fig. 7. Investment process generated by a well-trained policy network. *x*-axes represent the candidate nodes and *y*-axes represent the ADN management technologies. The dark block represents 1 and the light block represents 0.

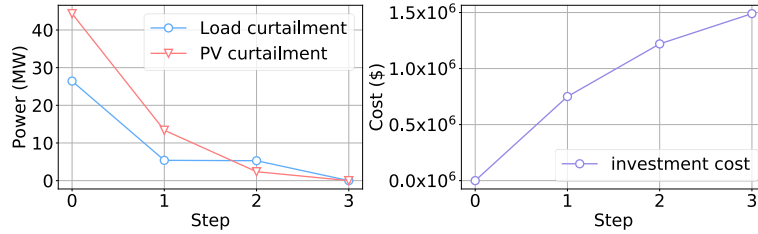


Fig. 8. The load and PV curtailment and the cost changing during each step of investment. The *x*-axes indicate the step of investment. The left *y*-axis indicates the power and right *y*-axis indicates the cost.

The investment in each technology at each node is presented in Fig. 7. At the last step, the required performances g^{expected} are satisfied. The nodes 14, 22, and 33 are equipped with ESS, gas generator, and CB but SVC is not selected during the planning process. In Fig. 8, the amounts of loss of load and PV curtailment decrease from 26.43MW and 44.34MW to 0.001MW and 0.003MW, respectively. The total investment cost is \$1,490,000. The operational results of each device are presented in Fig. 9-Fig. 13.

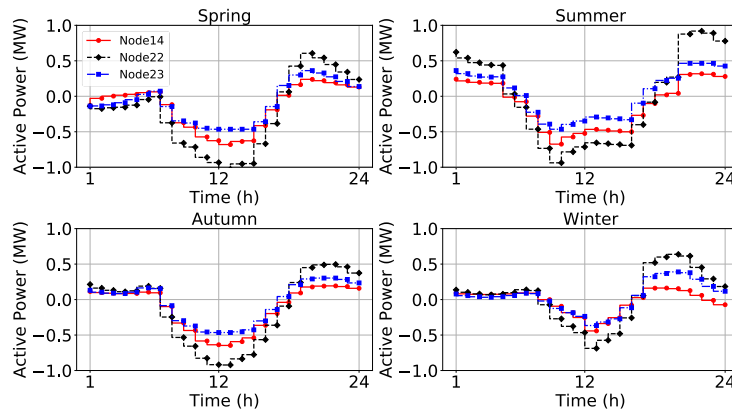


Fig. 9. Sequential active power output of the ESS. *x*-axes represent the time intervals and *y*-axes represent the output of ESS.

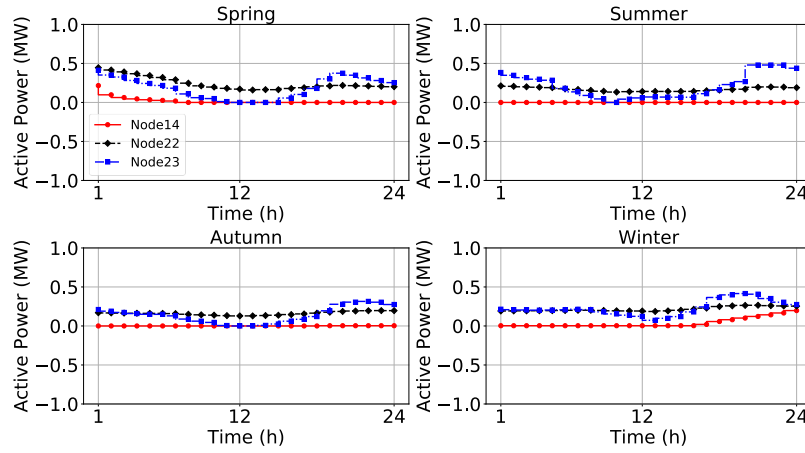


Fig. 10. Sequential active power output of the gas generator. *x*-axes represent the time intervals and *y*-axes represent the output of gas generators.

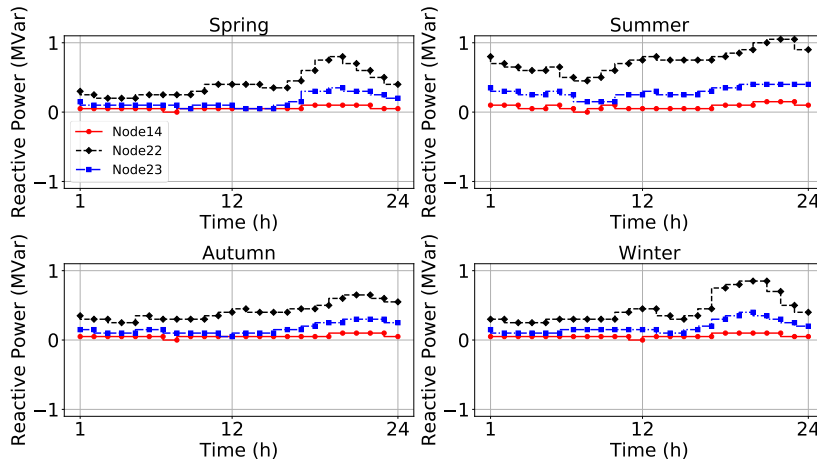


Fig. 11. Sequential reactive power compensation of CB. *x*-axes represent the time intervals and *y*-axes represent the output of the capacity banks.

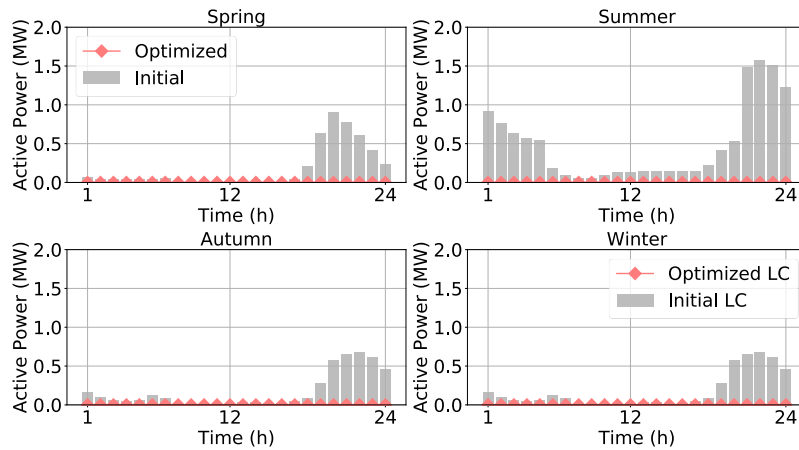


Fig. 12. Comparison of load curtailment. *x*-axes represent the time intervals and *y*-axes represent the amount of load curtailment.

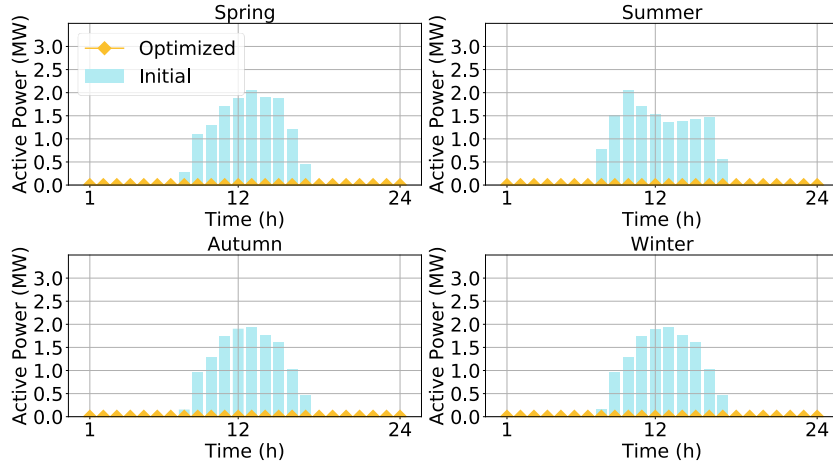


Fig. 13. Comparison of the PV curtailment. *x*-axes represent the time intervals and *y*-axes represent the PV curtailment.

In Fig. 9, all the ESSs remain charging to absorb the surplus power generated by PV during the daytime while release the energy to the grid during the peak demand period. In Fig. 10 and Fig. 11, the outputs of the gas generators and CBs increase during the peak demand period to avoid the loss of load and decrease when PV outputs ramp up. Hence, the integration of the RES and the load curtailment can significantly improve as presented in Fig. 12 and Fig. 13.

4.3. Evaluation under Various Cases

To show the advantages of the proposed method, the proposed case is compared with other cases as presented in Table 2.

Case	Planning decisions	Investment cost (10^4 \$)	LC (MWh)	PVC (MWh)
Initial	--	--	26.43	44.34
Case I	ESS(16,33) SVC(31) CB(5,15) ESS(1,7,14)	158	8.96	7.59
Case II	Gas(7,18,20,22,24) SVC(14,16,23,29) CB(7,14,18,20) ESS(1,7,13,16,33)	157	0.01	7.95
Case III	Gas(7,17) SVC(8,16,23) CB(5,15,23,30) ESS(14,22,33)	249	1.74	9.24
This paper	Gas(14,22,33) SVC(None) CB(14,22,33)	149	0.001	0.003

In Table 2, the numbers within the parenthesis indicate the candidate nodes for installing devices. The case I refers to the literature [11]. This case only explores a small part of the feasible space of the decision variables. Hence, the improvement of the performance is very limited although the investment cost is relatively small. Given that the cost of installing ESS is relatively high, we expand the candidate nodes of gas, SVC, and CB in the case II. The results show that the load curtailment is dramatically reduced but the PV curtailment still fails to reach the expected value. In the case III, we especially expand the set of candidate nodes for installing the ESS. However, the PV curtailment

dose not reduce but the investment cost increases. Compared with these four cases, the proposed method is capable of improving the network performance to the satisfied level at a lower investment cost.

4.4. Discussion

Current distribution network planning paradigm generally optimizes an objective function over a feasible region which is composed of several security constraints to obtain the optimal locations and sizes of the electrical devices. However, the desired performances may not be included in the proposed feasible region. Unlike the conventional planning method that optimizes a model to obtain the optimal location, size, and combination of electrical devices, this paper implements the MCTS-RL to dynamically update the planning scheme. Simulation results that the proposed method can produce a planning scheme that satisfies the desired performances i.e., both the load curtailment and PV curtailment less than 0.1MWh with a lower investment cost. As shown in Table 2, the proposed method achieves 0.001MWh of load curtailment and 0.003MWh of PV curtailment. Compared with the initial cases, the proposed method reduces the load curtailment and PV curtailment by 99.996% and 99.993%, respectively, at the cost of \$1,490,000. Case I uses the optimization model to determine the location and size of ESS, SVC, and capacity banks. The load curtailment and PV curtailment reduce by 66.10% and 82.88%, respectively. The investment cost is \$1,580,000. In case II, the gas generators are incorporated into the planning scheme. The load curtailment and PV curtailment reduce by 99.96% and 82.07%, respectively. The investment cost is \$1,570,000. In case III, the load curtailment and PV curtailment reduce by 93.41% and 79.16%, respectively. The investment cost is \$2,490,000. As shown in the numerical results, it can be seen that the model-based approaches fail to produce a planning scheme that satisfies the desired performances and cause a higher investment cost compared to the proposed method. Besides, the model-based approaches search for global optimum solutions over a fixed space. Once the global optimum is obtained, the yielded performance cannot be further improved. As shown in case I, although the obtained solution is a global optimum solution, there is still 8.65MWh PV curtailment and 7.60MWh load curtailment which could be further reduced.

5. Conclusion

This paper proposes a novel performance-oriented ADN planning method using the MCTS-RL. Instead of optimizing an objective function over a pre-determined solution space, the MCTS-RL is applied to dynamically update the APS until the expected performances are obtained. Numerical results show that the proposed method reduces the load curtailment and PV curtailment by 99.996% and 99.993%, respectively, and successfully achieves the desired performances at a low investment cost. As the future work, more devices such as feeders, charging facilities, and switches can be integrated into the element set and more complex performance indexes such as the reliability, social warfare, and EV hosting capacity can be adopted.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (5197133).

References

- [1] D. Rupolo, B. R. Pereira Junior, J. Contreras et al., "A new parallel and decomposition approach to solve the medium- and low-voltage planning of large-scale power distribution systems," *Int J Electr Power Energy Syst*, vol. 132, pp. 107191, 2021/11/01/, 2021.
- [2] M. Nick, R. Cherkaoui, and M. Paolone, "Optimal Planning of Distributed Energy Storage Systems in Active Distribution Networks Embedding Grid Reconfiguration," *IEEE Trans Power Syst*, vol. 33, no. 2, pp. 1577–1590, Mar, 2018.
- [3] A. H. Alobaidi, M. Khodayar, A. Vafamehr et al., "Stochastic expansion planning of battery energy storage for the interconnected distribution and data networks," *Int J Electr Power Energy Syst*, vol. 133, pp. 107231, 2021/12/01/, 2021.

- [4] X. Xu, J. Y. Li, Z. Xu *et al.*, “Enhancing photovoltaic hosting capacity-A stochastic approach to optimal planning of static var compensator devices in distribution networks,” *Appl Energy*, vol. 238, pp. 952-962, Mar 15, 2019.
- [5] Y. Amrane, M. Boudour, and M. Belazzoug, “A new Optimal reactive power planning based on Differential Search Algorithm,” *Int J Electr Power Energy Syst*, vol. 64, pp. 551-561, 2015/01/01/, 2015.
- [6] X. Shen, M. Shahidehpour, S. Zhu *et al.*, “Multi-Stage Planning of Active Distribution Networks Considering the Co-Optimization of Operation Strategies,” *IEEE Trans Smart Grid*, vol. 9, no. 2, pp. 1425-1433, 2018.
- [7] A. Barin, L. F. Pozzatti, L. N. Canha *et al.*, “Multi-objective analysis of impacts of distributed generation placement on the operational characteristics of networks for distribution system planning,” *Int J Electr Power Energy Syst*, vol. 32, no. 10, pp. 1157-1164, 2010/12/01/, 2010.
- [8] F. Ugranlı, “Analysis of renewable generation’s integration using multi-objective fashion for multistage distribution network expansion planning,” *Int J Electr Power Energy Syst*, vol. 106, pp. 301-310, 2019/03/01/, 2019.
- [9] M. Pehl, A. Arvesen, F. Humpenöder *et al.*, “Understanding future emissions from low-carbon power systems by integration of life-cycle assessment and integrated energy modelling,” *Nat Energy*, vol. 2, no. 12, pp. 939-945, 2017/12/01/, 2017.
- [10] T. D. de Lima, A. Tabares, N. Bañol Arias *et al.*, “Investment & generation costs vs CO2 emissions in the distribution system expansion planning: A multi-objective stochastic programming approach,” *Int J Electr Power Energy Syst*, vol. 131, pp. 106925, 2021/10/01/, 2021.
- [11] H. J. Gao, L. F. Wang, J. Y. Liu *et al.*, “Integrated Day-Ahead Scheduling Considering Active Management in Future Smart Distribution System,” *IEEE Trans Power Syst*, vol. 33, no. 6, pp. 6049-6061, Nov, 2018.
- [12] S. Xie, Z. Hu, L. Yang *et al.*, “Expansion planning of active distribution system considering multiple active network managements and the optimal load-shedding direction,” *Int J Electr Power Energy Syst*, vol. 115, pp. 105451, 2020/02/01/, 2020.
- [13] S. Gill, I. Koccar, and G. W. Ault, “Dynamic Optimal Power Flow for Active Distribution Networks,” *IEEE Trans Power Syst*, vol. 29, no. 1, pp. 121-131, 2014.
- [14] X. Q. Bai, L. Y. Qu, and W. Qiao, “Robust AC Optimal Power Flow for Power Networks With Wind Power Generation,” *IEEE Trans Power Syst*, vol. 31, no. 5, pp. 4163-4164, Sep, 2016.
- [15] S. F. Santos, D. Z. Fitiwi, M. Shafie-Khah *et al.*, “New Multistage and Stochastic Mathematical Model for Maximizing RES Hosting Capacity- Part I: Problem Formulation,” *IEEE Trans Sustain Energy*, vol. 8, no. 1, pp. 304-319, Jan, 2017.
- [16] A. Narimani, G. Nourbakhsh, A. Arefi *et al.*, “SAIDI Constrained Economic Planning and Utilization of Central Storage in Rural Distribution Networks,” *IEEE Syst J*, vol. 13, no. 1, pp. 842-853, Mar, 2019.
- [17] S. Heidari, M. Fotuhi-Firuzabad, and S. Kazemi, “Power Distribution Network Expansion Planning Considering Distribution Automation,” *IEEE Trans Power Syst*, vol. 30, no. 3, pp. 1261-1269, May, 2015.
- [18] Z. Ghofrani-Jahromi, M. Kazemi, and M. Ehsan, “Distribution Switches Upgrade for Loss Reduction and Reliability Improvement,” *IEEE Trans Power Deliv.*, vol. 30, no. 2, pp. 684-692, Apr, 2015.
- [19] A. Ehsan, and Q. Yang, “Coordinated Investment Planning of Distributed Multi-Type Stochastic Generation and Battery Storage in Active Distribution Networks,” *IEEE Trans Sustain Energy*, vol. 10, no. 4, pp. 1813-1822, Oct, 2019.
- [20] S. Mohtashami, D. Pudjianto, and G. Strbac, “Strategic Distribution Network Planning With Smart Grid Technologies,” *IEEE Trans Smart Grid*, vol. 8, no. 6, pp. 2656-2664, Nov, 2017.
- [21] M. Ghasemi, A. Kazemi, E. Bompard *et al.*, “A two-stage resilience improvement planning for power distribution systems against hurricanes,” *Int J Electr Power Energy Syst*, vol. 132, pp. 107214, 2021/11/01/, 2021.
- [22] A. Najafi Tari, M. S. Sepasian, and M. Tourandaz Kenari, “Resilience assessment and improvement of distribution networks against extreme weather events,” *Int J Electr Power Energy Syst*, vol. 125, pp. 106414, 2021/02/01/, 2021.
- [23] M. Mozaffari, H. A. Abyaneh, M. Jooshaki *et al.*, “Joint Expansion Planning Studies of EV Parking Lots Placement and Distribution Network,” *IEEE Trans Ind Inform*, vol. 16, no. 10, pp. 6455-6465, Oct., 2020.
- [24] O. D. Melgar-Dominguez, M. Pourakbari-Kasmaei, M. Lehtonen *et al.*, “An economic-environmental asset planning in electric distribution networks considering carbon emission trading and demand response,” *Electr Power Syst Res*, vol. 181, Apr, 2020.
- [25] X. Sun, and J. Qiu, “Two-Stage Volt/Var Control in Active Distribution Networks with Multi-Agent Deep Reinforcement Learning Method,” *IEEE Trans Smart Grid*, pp. 1-1, 2021.
- [26] H. Liu, and W. Wu, “Two-stage Deep Reinforcement Learning for Inverter-based Volt-VAR Control in Active Distribution Networks,” *IEEE Trans Smart Grid*, pp. 1-1, 2020.
- [27] Q. Yang, G. Wang, A. Sadeghi *et al.*, “Two-Timescale Voltage Control in Distribution Grids Using Deep Reinforcement Learning,” *IEEE Trans Smart Grid*, vol. 11, no. 3, pp. 2313-2323, 2020.
- [28] W. Wang, N. Yu, Y. Gao *et al.*, “Safe Off-Policy Deep Reinforcement Learning Algorithm for Volt-VAR Control in Power Distribution Systems,” *IEEE Trans Smart Grid*, vol. 11, no. 4, pp. 3008-3018, 2020.
- [29] Y. Gao, W. Wang, J. Shi *et al.*, “Batch-Constrained Reinforcement Learning for Dynamic Distribution Network Reconfiguration,” *IEEE Trans Smart Grid*, vol. 11, no. 6, pp. 5357-5369, 2020.
- [30] W. Hua, M. You and H. Sun, “Real-Time Price Elasticity Reinforcement Learning for Low Carbon Energy Hub Scheduling Based on Conditional Random Field,” 2019 IEEE/CIC International Conference on Communications Workshops in China (ICCC Workshops), 2019, pp. 204-209, doi: 10.1109/ICCCChinaW.2019.8849941.
- [31] M. Al-Saffar, and P. Musilek, “Reinforcement Learning-Based Distributed BESS Management for Mitigating Overvoltage Issues in Systems With High PV Penetration,” *IEEE Trans Smart Grid*, vol. 11, no. 4, pp. 2980-2994, Jul, 2020.
- [32] Y. W. Shang, W. C. Wu, J. W. Liao *et al.*, “Stochastic Maintenance Schedules of Active Distribution Networks Based on Monte-Carlo Tree Search,” *IEEE Trans Power Syst*, vol. 35, no. 5, pp. 3940-3952, Sept, 2020.
- [33] H. J. Gao, J. Y. Liu, and L. F. Wang, “Robust Coordinated Optimization of Active and Reactive Power in Active Distribution Systems,” *IEEE Trans Smart Grid*, vol. 9, no. 5, pp. 4436-4447, Sep, 2018.
- [34] D. Silver, T. Hubert, J. Schrittwieser *et al.*, “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play,” *Science*, vol. 362, no. 6419, pp. 1140-1144, Dec 7, 2018.
- [35] D. Silver, J. Schrittwieser, K. Simonyan *et al.*, “Mastering the game of Go without human knowledge,” *Nature*, vol. 550, no. 7676, pp. 354-359, Oct 19, 2017.