Dynamic Distribution Network Reconfiguration with Generation and Load Uncertainty

Shahab Bahrami *Member, IEEE*, Yu Christine Chen, *Member, IEEE*, and Vincent W. S. Wong, *Fellow, IEEE*

***Abstract*—Given the uncertainty in load demand and renewable energy sources, the distribution network reconfiguration (DNR) problem is a stochastic mixed-integer nonlinear optimization program with a running time that scales exponentially with the number of sectional and tie line switches. Stochastic optimization techniques require knowledge of the stochastic processes of the uncertain parameters, which may not be available in practice. This paper addresses both the scalability and uncertainty issues in solving the DNR problem by developing a deep reinforcement learning (DRL) algorithm that determines the optimal topology using a transformer deep neural network (DNN) architecture, and subsequently solves an AC optimal power flow (OPF) problem to satisfy the operation constraints. A neural combinatorial optimization algorithm is applied to train the DNN, which penalizes infeasible solutions. Simulations on a 119-bus test system show that our proposed algorithm can obtain a near-optimal solution to the stochastic DNR problem with a small gap (i.e., 4.7% on average) from the objective value of the deterministic DNR problem. When compared with existing learning-based DNR algorithms in the literature, our proposed algorithm can obtain at least 11% lower objective value. We demonstrate the scalability of our proposed algorithm in larger systems with 595, 1190, and 3570 buses.**

***Keywords*: deep reinforcement learning, distribution network reconfiguration, neural combinatorial optimization algorithm, optimal power flow, transformer deep neural network.**

I. Introduction

Dynamic reconfiguration of distribution networks is an effective approach for distribution network operators (DNOs) to enable network operation to be more resilient, especially during times with greater power demand, intermittent power supply, or transmission line failures. The solution of the distribution network reconfiguration (DNR) problem determines the status of the sectional and tie line switches to change the network topology. The DNR problem is typically formulated as a mixed-integer optimization program with the objective of minimizing the power losses, switching equipment wear-and-tear, cost of load interruption, and operation cost of backup generators, subject to the network operating constraints [1].

Remote-control switches play a crucial role in the reconfiguration of distribution networks. Although present-day distribution networks are still far from the extensive remote-control switching capability considered in this paper, industry efforts demonstrate a noticeable upward trend towards promoting and developing remote-control solutions for the distribution network automation. For example, Siemens [2] has developed a suite of distribution automation products that enable remote monitoring and control of switches with advanced features such as high-speed fault isolation and voltage regulation. Another example is the SCADA-mate switching system developed by S&C Electric Company [3], which allows for remote operation of the distribution networks to enhance efficiency and controllability. Other contributors such as Schneider Electric [4], Hubbell Power Systems [5], and G&W Electric [6] are proposing remote-control solutions to make the grid more adaptable and resilient. Moreover, Switched Source [7], has proposed advanced switching technologies to enhance distribution networks flexibility, making the system restoration faster and automated. Additionally, companies such as Allied Power and Control [8] and Caterpillar [9] are broadening the applications of remote-control switches with a range of solutions from panel-mounted devices to advanced switchgear systems. These industry efforts suggest growing recognition of long-term advantages of such investments to justify the initial costs of installing a large number of remote-control switches. Moreover, they signal broader adoption of remote-control switches useful for DNR in the near future.

The uncertainty in renewable energy sources and price responsive loads becomes more difficult to mitigate when the topology reconfiguration actions affect the operation of the distribution network in future time intervals due to inter-temporal operating constraints (e.g., constraint on the number of switching actions, ramp constraint for the backup diesel generators [10]). Furthermore, to guarantee a feasible network topology solution that satisfies system operating constraints, the DNR problem should embed the nonlinear AC power flow equations as constraints [11]–[13]. The consideration for the AC power flow equations and the uncertainty in generation and load render the DNR problem to be a stochastic mixed-integer nonlinear program (MINLP), which is difficult to solve.

There have been a number of efforts to solve the DNR problem with uncertainty in generation and load demand. We divide our review of the related work into two main categories. The first line of research focuses on applying optimization techniques such as approximate dynamic programming [14], stochastic optimization [15], [16], robust optimization [17], [18], second-order cone programming (SOCP) [19], [20], model predictive control [21], branch-and-bound algorithm [22], convex relaxation [23], and heuristic methods [24]–[29] to solve the DNR problem with uncertainty in generation and load demand. However, the aforementioned techniques require stochastic models of uncertain parameters. Furthermore, due to the computational complexity of the combinatorial optimization problems, it is a challenge to apply these techniques in

The authors are with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada, V6T 1Z4 (email: bahramis@ece.ubc.ca; chen@ece.ubc.ca; vincentw@ece.ubc.ca).

large-scale distribution systems with many buses and switches.

The second line of work addresses the potential limitations of optimization-based methods by applying learning-based algorithms to solve the DNR problem. Li *et al.* in [30] applied reinforcement learning to solve a multi-objective DNR problem, but the uncertainty in the generation and load demand was not considered. Huang *et al.* in [31] deployed deep convolutional neural networks to select a near-optimal network topology in the DNR problem with short-term voltage stability constraints. However, short-term planning does not guarantee long-term optimal operation of the distribution network with uncertainty in generation and load demand. To address long-term operation, Gao *et al.* in [32] applied deep reinforcement learning (DRL) with the actor-critic method to solve the DNR problem. Wang *et al.* in [33] developed a DNR algorithm using deep Q-learning with perturbations of the network weights to improve exploration during training. Zhang *et al.* in [34] studied DNR for service restoration in distribution networks and applied imitation learning to leverage prior information about the uncertain parameters. The proposed approaches in [32]–[34] are based on off-policy training method with historical data for power flow in different configuration scenarios of the distribution network, which may not be available in practice. Hence, the aforementioned algorithms may not guarantee a feasible power flow in the distribution network. To incorporate real-time variability of the uncertain parameters into the policy learning and decision-making, Li *et al.* in [35] proposed a safe DRL algorithm that hinges on constrained on-policy training method for the operation of the distribution networks. Moreover, Wang *et al.* in [36] developed a deep learning algorithm with on-policy approach to train a graph neural network that integrates the topology information of distribution networks into the learning and reconfiguration process. The algorithms proposed in [35] and [36] incorporate all network constraint violations as a penalty term in the objective function. While this strategy aids in penalizing infeasible solutions, the learning process can become highly sensitive to the penalty arising from constraint violations. This sensitivity can lead to diminished learning efficacy, particularly in large distribution networks, where understanding the feasibility of the action space in accordance with the network constraints is crucial for operation in practical distribution networks.

Unlike [32]–[34], in this paper, we take into account the distribution network constraints during the learning process to guarantee a feasible AC power flow solution to the DNR problem. Different from [35] and [36], our method integrates network constraints into an AC optimal power flow (OPF) problem, penalizing only infeasible binary solutions. This targeted approach facilitates the learning of feasible switch operations, load management, and backup generator deployment. Moreover, distinct from our previous work [37] that develops a fast optimization solver for the day-ahead unit commitment problem, the focus of this paper is online decision making for real-time distribution network reconfiguration. While [37] assumes a deterministic load forecast is available, the proposed DRL-based algorithm accounts for uncertain demand and renewable generation without explicit probabilistic or statistical model. The main contributions of this paper are as follows:

- *DRL-based Algorithm Design*: To tackle the uncertainty in the renewable generation and load demand, we formulate a Markov decision process (MDP) and develop a DRL algorithm with actor-critic method to solve the DNR problem. We use two deep neural networks (DNNs) corresponding to the policy and relative value function for the expected average cost, respectively. A multilayer perceptron DNN for the critic computes the relative value function. For the policy, we develop a transformer DNN architecture [38], [39], in order to obtain the set of optimal binary variables associated with the switches, interrupted loads, and operating backup diesel generators. We demonstrate that the solution to the relaxed version of the DNR problem can be used as an input for the policy network to accelerate the learning process. The trained actor and critic DNNs can be used to determine a near-optimal solution to the DNR problem in each time slot.
- *Feasibility of AC Power Flow Constraints*: It is difficult to obtain a policy that satisfies the AC power flow constraints. We decouple the tasks of obtaining a feasible power flow in the distribution network and obtaining binary variables. We apply a neural combinatorial optimization algorithm to train the policy DNN, which penalizes infeasible power flow solutions. To tackle the nonconvex AC power flow constraints, we apply convex relaxation techniques to transform the original DNR problem into an SOCP for a given binary solution. Thus, a single instance of SOCP can be solved in polynomial running time to obtain a feasible power flow, even in large-scale distribution networks.
- *Performance Evaluation*: Simulations are performed on a 119-bus distribution system to evaluate the performance of the proposed algorithm. Results show that the actor and critic DNNs can be trained in about 5000 iterations, which is acceptable for learning with historical sample data. We show that by performing multiple rounds of forward propagation, the trained policy DNN always obtains a feasible solution for the binary variables and power flow in the distribution network. Results also show that the solution obtained by our proposed algorithm is near-optimal with an average gap of 4.7% from the optimal solution of a deterministic DNR problem with complete information about future load demand and renewable generation in ten days. When compared with the algorithm in [32], the gap between the expected average cost with our proposed algorithm is 11% smaller. Also, our proposed algorithm outperforms the approaches in [35] and [36], which penalize all operation constraints, by achieving 24.5% lower expected average cost. We also demonstrate the scalability of our proposed algorithm by solving the stochastic DNR problem in large test systems with 595, 1190, and 3570 buses.

The remainder of this paper is organized as follows. In Section II, we formulate the DNR problem as an MDP. In Section III, we propose a DRL algorithm to solve the underlying MDP. Section IV evaluates the performance of the proposed algorithm via simulations. We conclude the paper in Section V.

II. MDP Formulation of the DNR Problem

Consider a distribution network consisting of a set of buses $\mathcal{N} = \{1, \ldots, N\}$ and a set of transmission lines $\mathcal{L} \subseteq \mathcal{N} \times \mathcal{N}$. Let $\mathcal{N}^{\text{sub}} \subset \mathcal{N}$ denote the set of substation buses. We use $\mathcal{N}^{-} = \mathcal{N} \setminus \mathcal{N}^{\text{sub}}$ to denote the set of buses other than the substation buses. Bus $n \in \mathcal{N}^{-}$ may be connected to a backup diesel generator or renewable sources, e.g., photovoltaic (PV) systems or wind energy conversion systems. The network reconfiguration can be performed as part of a long-term plan with look-ahead flexibility in anticipating operational issues, e.g., power supply shortage in substation buses and failure in branches. Hence, we consider the DNR problem with an infinite operation horizon denoted by a set $\mathcal{T} = \{1, 2, \ldots\}$ of time slots with equal duration Δt (e.g., 15 minutes). The inherent variability of renewable energy sources, combined with the unpredictable nature of load demands, can affect the network's operational dynamics. Thus, by integrating these sources of uncertainty, our approach aims to provide a more practical and adaptive solution for the DNR problem. We formulate the DNR problem as an MDP with expected average cost [40, Ch. 3] because the state of the distribution network in the next time slot depends only on the state in the current time slot and the operation action taken. This study introduces a comprehensive framework for DNR that incorporates not only switch operations but also the deployment of backup generators and the management of flexible loads during demand peaks or power deficit. The proposed approach is aligned with real-world operational goals of DNOs to maintain secure and resilient distribution systems. We develop a DRL algorithm that learns to guarantee the inter-temporal constraints and results in a near-optimal solution while satisfying power flow constraints. Next, we describe the system state, action, operating constraints, policy, cost, and relative value function.

A. System State

Variables pertinent to the feasibility of network reconfiguration in time slot $t \in \mathcal{T}$ include the state for the switches, loads, backup generators, renewable sources, and supplied power from substation buses. The DNO determines the topology of the distribution network in time slot t by setting the binary variable $\alpha_{mn,t} \in \{0,\, 1\}$ to close (i.e., $\alpha_{mn,t} = 1$) or open (i.e., $\alpha_{mn,t} = 0$) the switch on line $(m,n) \in \mathcal{L}$. In order to mitigate the practical wear-and-tear cost of closing or opening the switches, we impose the limit t^{sw} for the number of time slots between two consecutive status changes for a switch on line $(m,n) \in \mathcal{L}$. We define the binary variable $\lambda_{mn,t} \in \{0,\, 1\}$ to indicate whether the status of a switch on line $(m,n) \in \mathcal{L}$ is changed in time slot t (i.e., $\lambda_{mn,t} = 1$) or not (i.e., $\lambda_{mn,t} = 0$). We also define the auxiliary variable $\Lambda_{mn,t} \in [0,\, 1]$ as the indicator that whether the status of a switch on line $(m,n) \in \mathcal{L}$ is changed in the previous $t^{\text{sw}} - 1$ time slots (i.e., $\Lambda_{mn,t} > 0$) or not (i.e., $\Lambda_{mn,t} = 0$). As part of the control action in time slot t, binary variable $d_{n,t} \in \{0,\, 1\}$ indicates whether to serve ($d_{n,t} = 1$) or disconnect ($d_{n,t} = 0$) the load at bus $n \in \mathcal{N}^{-}$. Let $P_{D_{n,t}}$ and $Q_{D_{n,t}}$, respectively, denote the active-power and reactive-power components of the load demand at bus n in time slot t. Backup diesel generators can be used in case of generation shortage. The binary variable $u_{n,t} \in \{0,\, 1\}$ indicates whether the diesel generator at bus $n \in \mathcal{N}^{-}$ is on ($u_{n,t} = 1$) or off ($u_{n,t} = 0$) in time slot t. Let $P_{G_{n,t}}$ and $Q_{G_{n,t}}$, respectively, denote the active-power and reactive-power outputs of the backup generator at bus n in time slot t. Let $P^{\text{r}}_{G_{n,t}}$ and $Q^{\text{r}}_{G_{n,t}}$, respectively, denote the active- and reactive-power outputs of the renewable sources at bus n in time slot t. Let $P^{\text{sub,max}}_{G_n,t}$ denote the upper limit for the active-power supply at the substation bus $n \in \mathcal{N}^{\text{sub}}$ in time t.

The system state in time slot $t \in \mathcal{T}$ includes the status of the switches at the end of previous time slot $t-1$, the on/off status and the setpoint of backup generators at the end of previous time slot $t-1$, the active- and reactive-power components of the load and the renewable generation, and the upper limit for the active-power supply at substation buses in current time slot t. We define system state in time slot t as $\boldsymbol{s}_t = ((\alpha_{mn,t-1},\, \lambda_{mn,t-1},\, \Lambda_{mn,t-1},\, (m,n) \in \mathcal{L}),\, (u_{n,t-1},\, P_{G_{n,t-1}},\, P_{D_{n,t}},\, Q_{D_{n,t}},\, P^{\text{r}}_{G_{n,t}},\, n \in \mathcal{N}^{-}),\, (P^{\text{sub,max}}_{G_n,t},\, n \in \mathcal{N}^{\text{sub}}))$. We use $\mathcal{S}$ to denote the set of system states. Let $x(\boldsymbol{s}_t)$ and $\boldsymbol{x}(\boldsymbol{s}_t)$, respectively, denote the values taken by scalar and vector variables x and $\boldsymbol{x}$ in state $\boldsymbol{s}_t \in \mathcal{S}$ in time slot $t \in \mathcal{T}$.

B. Action and Operating Constraints

The action in state $\boldsymbol{s}_t \in \mathcal{S}$ includes the decision variables to control the switches, backup generators, loads, and power flow in the distribution network. Given $\boldsymbol{s}_t \in \mathcal{S}$, the binary decision variables include

$$\alpha_{mn}(\boldsymbol{s}_t),\, u_n(\boldsymbol{s}_t),\, d_n(\boldsymbol{s}_t) \in \{0,1\},\, (m,n) \in \mathcal{L},\, n \in \mathcal{N}^{-}. \quad (1)$$

We assume that the DNO has complete knowledge regarding the network's topology. This assumption allows us to make informed decisions about reconfiguration, ensuring that the resulting network configuration is both feasible and optimized for the given operational conditions. The distribution network typically operates in a radial topology and must remain connected. That is, by considering an arbitrary bus as the root bus, any subset of buses should not be isolated from the root. We set bus N to be the root bus. We define binary variable $\beta_{mn}(\boldsymbol{s}_t)$, $(m,n) \in \mathcal{L}$ and use the following spanning tree constraints in state $\boldsymbol{s}_t \in \mathcal{S}$:

$$\sum_{(m,n)\in\mathcal{L}} \alpha_{mn}(\boldsymbol{s}_t) = N - 1, \quad (2a)$$

$$\sum_{(m,n)\in\mathcal{L}} \beta_{mn}(\boldsymbol{s}_t) - \sum_{(n,k)\in\mathcal{L}} \beta_{nk}(\boldsymbol{s}_t) = 1, \;\; n \in \mathcal{N} \setminus \{N\}, \quad (2b)$$

$$\sum_{(m,N)\in\mathcal{L}} \beta_{mN}(\boldsymbol{s}_t) - \sum_{(N,k)\in\mathcal{L}} \beta_{Nk}(\boldsymbol{s}_t) = -N + 1, \quad (2c)$$

$$0 \leq \beta_{mn}(\boldsymbol{s}_t) \leq \alpha_{mn}(\boldsymbol{s}_t), \qquad (m,n) \in \mathcal{L}. \quad (2d)$$

Constraint (2a) is a necessary condition for a radial topology for the distribution system, ensuring that the network is connected but with no cycles. Constraints (2b)−(2d) guarantee that no subset of buses is isolated from reference bus N to maintain the connectivity of the distribution network.

The limit t^{sw} for the number of time slots between two consecutive status changes for a switch on line $(m,n) \in \mathcal{L}$ imposes an inter-temporal constraint $\sum_{t'=t-t^{\text{sw}}}^{t} \big|\alpha_{mn}(\boldsymbol{s}_{t'}) -$

$\alpha_{mn}(\boldsymbol{s}_{t'-1})\big| \leq 1$ for $t' > 1$. It implies that the switching action for a switch on line $(m,n) \in \mathcal{L}$ in the current time slot t depends on the switching actions in the previous t^{sw} time slots. To formulate the problem as an MDP, we perform some algebraic manipulations to obtain a set of inter-temproal constraints that only contain the decision variables for the current time slot t and the previous time slot $t-1$. If the value of the auxiliary variable $\Lambda_{mn}(\boldsymbol{s}_t) = 0$, then the status of a switch on line $(m,n) \in \mathcal{L}$ has not been changed in the previous $t^{\text{sw}} - 1$ time slots. Therefore, the status of the switch on line $(m,n) \in \mathcal{L}$ can be changed in time slot t. Otherwise, the status of the switch on line $(m,n) \in \mathcal{L}$ remains unchanged in time slot t. We have the following inter-temporal constraints for the switch on line $(m,n) \in \mathcal{L}$ in time slot $t \in \mathcal{T}$:

$$\begin{aligned}\Lambda_{mn}(\boldsymbol{s}_t) = \max\big\{0,\, (\Lambda_{mn}(\boldsymbol{s}_{t-1}) - \epsilon)(1 - \lambda_{mn}(\boldsymbol{s}_{t-1}))\big\}&\\ + \lambda_{mn}(\boldsymbol{s}_{t-1}),&\end{aligned} \tag{3a}$$

$$\lambda_{mn}(\boldsymbol{s}_t) + \Lambda_{mn}(\boldsymbol{s}_t) \leq 1, \tag{3b}$$

$$\lambda_{mn}(\boldsymbol{s}_t) = \big|\alpha_{mn}(\boldsymbol{s}_t) - \alpha_{mn}(\boldsymbol{s}_{t-1})\big|, \tag{3c}$$

where $\epsilon = 1/t^{\text{sw}}$ is a positive constant.

Let $P_{G_n}^{\text{min}}$ and $P_{G_n}^{\text{max}}$, respectively, denote the lower and upper limits for the active-power output of the diesel generator at bus $n \in \mathcal{N}^-$. We denote the maximum ramp-up and ramp-down rates for the diesel generator at bus n by r_n^{u} and r_n^{d}, respectively. For $n \in \mathcal{N}^-$ and $\boldsymbol{s}_t \in \mathcal{S}$, we have

$$u_n(\boldsymbol{s}_t) P_{G_n}^{\text{min}} \leq P_{G_n}(\boldsymbol{s}_t) \leq u_n(\boldsymbol{s}_t) P_{G_n}^{\text{max}}, \tag{4a}$$

$$P_{G_n}(\boldsymbol{s}_t) - P_{G_n}(\boldsymbol{s}_{t-1}) \leq u_n(\boldsymbol{s}_{t-1})\, r_n^{\text{u}}, \tag{4b}$$

$$P_{G_n}(\boldsymbol{s}_{t-1}) - P_{G_n}(\boldsymbol{s}_t) \leq u_n(\boldsymbol{s}_{t-1})\, r_n^{\text{d}}. \tag{4c}$$

As a synchronous generator, the loading capability of the diesel generator at bus n determines the limit for its reactive-power output. The loading capability is obtained by the limits for the generator armature current, field current, and under-excitation [41, Ch. 5]. For $n \in \mathcal{N}^-$ and $\boldsymbol{s}_t \in \mathcal{S}$, we have

$$\big(P_{G_n}(\boldsymbol{s}_t)\big)^2 + \big(Q_{G_n}(\boldsymbol{s}_t)\big)^2 \leq u_n(\boldsymbol{s}_t)\big(P_{G_n}^{\text{max}}(\boldsymbol{s}_t)\big)^2, \tag{5a}$$

$$\big(P_{G_n}(\boldsymbol{s}_t)\big)^2 + \big(Q_{G_n}(\boldsymbol{s}_t) - Q_{G_n}^{\text{f}}\big)^2 \leq \big(Q_{G_n}^{\text{max}} - Q_{G_n}^{\text{f}}\big)^2, \tag{5b}$$

$$\big(P_{G_n}(\boldsymbol{s}_t)\big)^2 + \big(Q_{G_n}(\boldsymbol{s}_t) - Q_{G_n}^{\text{e}}\big)^2 \leq \big(Q_{G_n}^{\text{e}} - Q_{G_n}^{\text{min}}\big)^2, \tag{5c}$$

where $Q_{G_n}^{\text{min}}$ and $Q_{G_n}^{\text{max}}$, respectively, denote the lower and upper limits for the reactive-power output of the diesel generator at bus n, and $Q_{G_n}^{\text{f}}$ and $Q_{G_n}^{\text{e}}$ are positive and negative constant parameters, respectively, which depend on the stator and rotor heat limits of the diesel generator at bus n.

The reactive-power output of the renewable source at bus n is limited by lower bound $Q_{G_n}^{\text{r,min}}$ and upper bound $Q_{G_n}^{\text{r,max}}$. For $n \in \mathcal{N}^-$ and $\boldsymbol{s}_t \in \mathcal{S}$, we have

$$Q_{G_n}^{\text{r,min}} \leq Q_{G_n}^{\text{r}}(\boldsymbol{s}_t) \leq Q_{G_n}^{\text{r,max}}. \tag{6}$$

The active-power supply at the substation bus $n \in \mathcal{N}^{\text{sub}}$ is limited by $P_{G_n}^{\text{sub,max}}(\boldsymbol{s}_t)$ in state $\boldsymbol{s}_t$. We have

$$0 \leq P_{G_n}^{\text{sub}}(\boldsymbol{s}_t) \leq P_{G_n}^{\text{sub,max}}(\boldsymbol{s}_t), \;\; \boldsymbol{s}_t \in \mathcal{S},\, n \in \mathcal{N}^{\text{sub}}. \tag{7}$$

Next, we describe the power flow constraints imposed by the distribution network. In our framework, we assume complete and accurate information concerning the parameters of network transmission lines and the admittance matrix. This detailed knowledge enables us to precisely compute power flows and losses, ensuring that the reconfiguration decisions adhere to the operational constraints of the distribution system. We use the lumped-element Π model for transmission lines [42]. Let r_{mn} and x_{mn}, respectively, denote the series resistance and reactance for line $(m,n) \in \mathcal{L}$. Let g_n and b_n, respectively, denote the shunt conductance and susceptance for bus $n \in \mathcal{N}$. Let $P_{mn}(\boldsymbol{s}_t)$ and $Q_{mn}(\boldsymbol{s}_t)$, respectively, denote the active- and reactive-power flow on line $(m,n) \in \mathcal{L}$ from bus m to bus n in state $\boldsymbol{s}_t$. Let $V_n(\boldsymbol{s}_t)$ denote the voltage phasor of bus n in state $\boldsymbol{s}_t$. We define variable $z_n(\boldsymbol{s}_t) = |V_n(\boldsymbol{s}_t)|^2$ as the squared voltage magnitude of bus n in state $\boldsymbol{s}_t$. Let $I_{mn}(\boldsymbol{s}_t)$ denote the current phasor for line (m,n) in state $\boldsymbol{s}_t$. We define variable $\ell_{mn}(\boldsymbol{s}_t) = |I_{mn}(\boldsymbol{s}_t)|^2$ as the squared current magnitude for line (m,n) in state $\boldsymbol{s}_t$. Let $Q_{G_n}^{\text{sub}}(\boldsymbol{s}_t)$ denote the reactive-power injected into the substation bus $n \in \mathcal{N}^{\text{sub}}$ in state $\boldsymbol{s}_t$. We have the following active- and reactive-power balance equations:

$$\begin{aligned}P_n^{\text{inj}}(\boldsymbol{s}_t) = \sum_{(n,k)\in\mathcal{L}} P_{nk}(\boldsymbol{s}_t) - \sum_{(m,n)\in\mathcal{L}} \Big(P_{mn}(\boldsymbol{s}_t)&\\ - r_{mn}\ell_{mn}(\boldsymbol{s}_t)\Big) + g_n z_n(\boldsymbol{s}_t), \quad n \in \mathcal{N},&\end{aligned} \tag{8a}$$

$$\begin{aligned}Q_n^{\text{inj}}(\boldsymbol{s}_t) = \sum_{(n,k)\in\mathcal{L}} Q_{nk}(\boldsymbol{s}_t) - \sum_{(m,n)\in\mathcal{L}} \Big(Q_{mn}(\boldsymbol{s}_t)&\\ - x_{mn}\ell_{mn}(\boldsymbol{s}_t)\Big) + b_n z_n(\boldsymbol{s}_t), \quad n \in \mathcal{N},&\end{aligned} \tag{8b}$$

where $P_n^{\text{inj}}(\boldsymbol{s}_t)$ and $Q_n^{\text{inj}}(\boldsymbol{s}_t)$, respectively, are the active- and reactive-power injected into bus $n \in \mathcal{N}$. For substation bus $n \in \mathcal{N}^{\text{sub}}$, we have $P_n^{\text{inj}}(\boldsymbol{s}_t) = P_{G_n}^{\text{sub}}(\boldsymbol{s}_t)$ and $Q_n^{\text{inj}}(\boldsymbol{s}_t) = Q_{G_n}^{\text{sub}}(\boldsymbol{s}_t)$. For bus $n \in \mathcal{N}^-$, we have $P_n^{\text{inj}}(\boldsymbol{s}_t) = P_{G_n}(\boldsymbol{s}_t) + P_{G_n}^{\text{r}}(\boldsymbol{s}_t) - d_n(\boldsymbol{s}_t)P_{D_n}(\boldsymbol{s}_t)$ and $Q_n^{\text{inj}}(\boldsymbol{s}_t) = Q_{G_n}(\boldsymbol{s}_t) + Q_{G_n}^{\text{r}}(\boldsymbol{s}_t) - d_n(\boldsymbol{s}_t)Q_{D_n}(\boldsymbol{s}_t)$.

Given system state $\boldsymbol{s}_t$, the difference of the squared voltage magnitudes for the buses on line $(m,n) \in \mathcal{L}$ is obtained as $z_m(\boldsymbol{s}_t) - z_n(\boldsymbol{s}_t) = (r_{mn}^2 + x_{mn}^2)\ell_{mn}(\boldsymbol{s}_t) - 2\big(r_{mn}P_{mn}(\boldsymbol{s}_t) + x_{mn}Q_{mn}(\boldsymbol{s}_t)\big)$. This equality constraint is equivalent to the following two inequality constraints for $(m,n) \in \mathcal{L}$:

$$\begin{aligned}z_m(\boldsymbol{s}_t) - z_n(\boldsymbol{s}_t) \leq M\big(1 - \alpha_{mn}(\boldsymbol{s}_t)\big) + (r_{mn}^2 + x_{mn}^2)\ell_{mn}(\boldsymbol{s}_t)&\\ - 2\big(r_{mn}P_{mn}(\boldsymbol{s}_t) + x_{mn}Q_{mn}(\boldsymbol{s}_t)\big),&\end{aligned} \tag{9a}$$

$$\begin{aligned}z_m(\boldsymbol{s}_t) - z_n(\boldsymbol{s}_t) \geq -M\big(1 - \alpha_{mn}(\boldsymbol{s}_t)\big) + (r_{mn}^2 + x_{mn}^2)\ell_{mn}(\boldsymbol{s}_t)&\\ - 2\big(r_{mn}P_{mn}(\boldsymbol{s}_t) + x_{mn}Q_{mn}(\boldsymbol{s}_t)\big),&\end{aligned} \tag{9b}$$

where M is a sufficiently large parameter (e.g., $M = 10^3$). The squared apparent power flow from bus m to bus n, $(m,n) \in \mathcal{L}$, in state $\boldsymbol{s}_t \in \mathcal{S}$, can be obtained as follows:

$$\ell_{mn}(\boldsymbol{s}_t)\, z_m(\boldsymbol{s}_t) = \big(P_{mn}(\boldsymbol{s}_t)\big)^2 + \big(Q_{mn}(\boldsymbol{s}_t)\big)^2. \tag{10}$$

We denote the lower and upper limits of the voltage magnitude at bus $n \in \mathcal{N}$ by V_n^{min} and V_n^{max}, respectively. Let I_{mn}^{max} denote the upper limit of the current magnitude in line (m,n). The following constraints express the limits on the voltage magnitude of buses and current magnitude of lines in $\boldsymbol{s}_t \in \mathcal{S}$:

$$(V_n^{\text{min}})^2 \leq z_n(\boldsymbol{s}_t) \leq (V_n^{\text{max}})^2, \qquad n \in \mathcal{N}, \tag{11a}$$

$$\ell_{mn}(\boldsymbol{s}_t) \leq \alpha_{mn}(\boldsymbol{s}_t)(I_{mn}^{\text{max}})^2, \qquad (m,n) \in \mathcal{L}. \tag{11b}$$

In summary, the action in state $\boldsymbol{s}_t \in \mathcal{S}$ is defined as vector $\boldsymbol{a}(\boldsymbol{s}_t) = ((\alpha_{mn}(\boldsymbol{s}_t),\, \beta_{mn}(\boldsymbol{s}_t),\, \lambda_{mn}(\boldsymbol{s}_t),\, Q_{mn}(\boldsymbol{s}_t),\, \ell_{mn}(\boldsymbol{s}_t),\, (m,n) \in \mathcal{L}),\, (u_n(\boldsymbol{s}_t),\, P_{G_n}(\boldsymbol{s}_t),\, Q_{G_n}(\boldsymbol{s}_t),\, Q^{\mathrm{r}}_{G_n}(\boldsymbol{s}_t),\, d_n(\boldsymbol{s}_t),\, n \in \mathcal{N}^-),\, (P^{\mathrm{sub}}_{G_n}(\boldsymbol{s}_t),\, Q^{\mathrm{sub}}_{G_n}(\boldsymbol{s}_t),\, n \in \mathcal{N}^{\mathrm{sub}}),\, (z_n(\boldsymbol{s}_t),\, n \in \mathcal{N}))$. Given state $\boldsymbol{s}_t \in \mathcal{S}$, let $\mathcal{A}(\boldsymbol{s}_t)$ denote the feasible action space defined by constraints (1)−(11).

C. Policy and Immediate Cost

We consider a stationary random policy as $\boldsymbol{\pi} = (\boldsymbol{\pi}(\boldsymbol{s}),\, \boldsymbol{s} \in \mathcal{S})$, where $\boldsymbol{\pi}(\boldsymbol{s}) = (\pi(\boldsymbol{a}(\boldsymbol{s}) \,|\, \boldsymbol{s}),\, \boldsymbol{a}(\boldsymbol{s}) \in \mathcal{A}(\boldsymbol{s}))$ specifies the probability $\pi(\boldsymbol{a}(\boldsymbol{s}) \,|\, \boldsymbol{s})$ of choosing a feasible action $\boldsymbol{a}(\boldsymbol{s}) \in \mathcal{A}(\boldsymbol{s})$ in a given state $\boldsymbol{s} \in \mathcal{S}$.

The objective of the DNR problem is to jointly minimize the network losses along with the costs of switching equipment wear-and-tear, operating the backup generators, and interrupting loads. The network losses in state $\boldsymbol{s}_t$ can be obtained as $\sum_{(m,n)\in\mathcal{L}} r_{mn}\ell_{mn}(\boldsymbol{s}_t)$. We scale the network losses by a weighting coefficient η^{loss} (in \$/kW) to convert into a monetary unit. The switching cost in state $\boldsymbol{s}_t$ can be obtained as $\sum_{(m,n)\in\mathcal{L}} \mu^{\mathrm{switch}}_{mn} \left|\alpha_{mn}(\boldsymbol{s}_t) - \alpha_{mn}(\boldsymbol{s}_{t-1})\right|$, where $\mu^{\mathrm{switch}}_{mn}$ is a nonnegative weighting coefficient in monetary unit that captures the cost incurred by closing or opening the switch on line $(m,n) \in \mathcal{L}$. The operation cost of a backup generator at bus n with generation level $P_{G_n}(\boldsymbol{s}_t)$ in state $\boldsymbol{s}_t$ can be modelled by a linear function $c_{n1}P_{G_n}(\boldsymbol{s}_t) + c_{n0}\, u_n(\boldsymbol{s}_t)$, where c_{n0} and c_{n1} are nonnegative coefficients [43]. The load interruption cost (referred to as the value of lost load (VoLL) [44, Ch.13]) in state $\boldsymbol{s}_t$ can be computed as $\sum_{n\in\mathcal{N}^-}(1 - d_n(\boldsymbol{s}_t))\,\omega^{\mathrm{load}}_n P_{D_n}(\boldsymbol{s}_t)$, where the nonnegative weighting coefficient ω^{load}_n (in \$/kW) captures the interruption cost of the load at bus n. A critical must-run load at bus n would have a large value of ω^{load}_n, whereas a flexible load at bus n would have a small value of ω^{load}_n. The immediate cost with action $\boldsymbol{a}(\boldsymbol{s}_t)$ in state $\boldsymbol{s}_t$ is obtained as follows:

$$\begin{aligned} c(\boldsymbol{s}_t, \boldsymbol{a}(\boldsymbol{s}_t)) = &\sum_{(m,n)\in\mathcal{L}} \Big(\eta^{\mathrm{loss}}\, r_{mn}\ell_{mn}(\boldsymbol{s}_t) \\ &+ \mu^{\mathrm{switch}}_{mn}\big|\alpha_{mn}(\boldsymbol{s}_t) - \alpha_{mn}(\boldsymbol{s}_{t-1})\big|\Big) + \sum_{n\in\mathcal{N}^-}\Big(c_{n1}P_{G_n}(\boldsymbol{s}_t) \\ &+ c_{n0}\, u_n(\boldsymbol{s}_t) + \big(1 - d_n(\boldsymbol{s}_t)\big)\,\omega^{\mathrm{load}}_n P_{D_n}(\boldsymbol{s}_t)\Big). \end{aligned} \tag{12}$$

Via weighting factors η^{loss}, $\mu^{\mathrm{switch}}_{mn}$, $(m,n) \in \mathcal{L}$, and ω^{load}_n, $n \in \mathcal{N}^-$, the immediate cost in (12) offers a balance between the operational needs of the distribution network and the associated cost.

D. Relative Value Function

We define the expected average cost for a given policy $\boldsymbol{\pi}$ as follows:

$$\rho^{\boldsymbol{\pi}} = \lim_{T\to\infty} \mathbb{E}^{\boldsymbol{\pi}}\left\{\frac{1}{T}\, \sum_{t=1}^{T} c(\boldsymbol{s}_t, \boldsymbol{a}(\boldsymbol{s}_t))\right\}, \tag{13}$$

where $\mathbb{E}^{\boldsymbol{\pi}}\{\cdot\}$ is the expectation over selecting feasible actions for the given policy $\boldsymbol{\pi}$. The expected average cost in (13) is bounded and does not depend on the initial state if we assume that the process is ergodic, i.e., the probability of reaching any state from any other is nonzero [40, Ch. 11]. For a given policy $\boldsymbol{\pi}$, we define the relative value function as follows:

$$V^{\boldsymbol{\pi}}(\boldsymbol{s}) = \sum_{t'=0}^{\infty} \mathbb{E}^{\boldsymbol{\pi}}\Big\{c(\boldsymbol{s}_{t+t'}, \boldsymbol{a}(\boldsymbol{s}_{t+t'})) - \rho^{\boldsymbol{\pi}} \;\Big|\; \boldsymbol{s}_t = \boldsymbol{s}\Big\}. \tag{14}$$

We define the transition probability $\Pr(\boldsymbol{s}' \,|\, \boldsymbol{s},\, \boldsymbol{a}(\boldsymbol{s}))$ from state $\boldsymbol{s}$ to $\boldsymbol{s}'$ with action $\boldsymbol{a}(\boldsymbol{s})$. The DNO aims to obtain policy $\boldsymbol{\pi}$ such that the relative value function is minimized over all states $\boldsymbol{s} \in \mathcal{S}$. This is equivalent to solving the following Bellman equations:

$$\begin{aligned} \boldsymbol{\mathcal{P}}^{\mathrm{MDP}}: \quad V^{\boldsymbol{\pi}}(\boldsymbol{s}) + \rho^{\boldsymbol{\pi}} = \underset{\boldsymbol{a}(\boldsymbol{s})\in\mathcal{A}(\boldsymbol{s})}{\text{minimize}}\; &\Big\{c(\boldsymbol{s}, \boldsymbol{a}(\boldsymbol{s})) + \\ &\sum_{\boldsymbol{s}'\in\mathcal{S}} \Pr\big(\boldsymbol{s}' \,|\, \boldsymbol{s},\, \boldsymbol{a}(\boldsymbol{s})\big)\, V^{\boldsymbol{\pi}}(\boldsymbol{s}')\Big\}, \quad \forall\, \boldsymbol{s} \in \mathcal{S}. \end{aligned}$$

Obtaining an optimal policy that solves problem $\boldsymbol{\mathcal{P}}^{\mathrm{MDP}}$ is challenging. The action $\boldsymbol{a}(\boldsymbol{s}_t)$ obtained by policy $\boldsymbol{\pi}(\boldsymbol{s}_t)$ is a combination of binary variables and continuous variables for power flow in the distribution network. Moreover, the feasible action space $\mathcal{A}(\boldsymbol{s}_t)$ includes constraints with binary variables and the nonconvex constraint (10) for the apparent power flow in distribution lines. To address this challenge, we divide the action vector $\boldsymbol{a}(\boldsymbol{s}_t)$ into vectors $\boldsymbol{\phi}(\boldsymbol{s}_t) = ((\alpha_{mn}(\boldsymbol{s}_t),\, (m,n) \in \mathcal{L}),\, (u_n(\boldsymbol{s}_t),\, d_n(\boldsymbol{s}_t),\, n \in \mathcal{N}^-))$ consisting of binary variables and $\boldsymbol{\psi}(\boldsymbol{s}_t) = ((\beta_{mn}(\boldsymbol{s}_t),\, \lambda_{mn}(\boldsymbol{s}_t),\, Q_{mn}(\boldsymbol{s}_t),\, \ell_{mn}(\boldsymbol{s}_t),\, (m,n) \in \mathcal{L}),\, (P_{G_n}(\boldsymbol{s}_t),\, Q_{G_n}(\boldsymbol{s}_t),\, Q^{\mathrm{r}}_{G_n}(\boldsymbol{s}_t),\, n \in \mathcal{N}^-),\, (P^{\mathrm{sub}}_{G_n}(\boldsymbol{s}_t),\, Q^{\mathrm{sub}}_{G_n}(\boldsymbol{s}_t),\, n \in \mathcal{N}^{\mathrm{sub}}),\, (z_n(\boldsymbol{s}_t),\, n \in \mathcal{N}))$ consisting of continuous variables. We decouple the tasks of (i) obtaining the action vector $\boldsymbol{\phi}(\boldsymbol{s}_t)$ for the status of the switches, backup generators, and loads, and (ii) obtaining the action vector $\boldsymbol{\psi}(\boldsymbol{s}_t)$ for a feasible power flow in the distribution network. Furthermore, we rewrite the immediate cost (12) as follows:

$$c(\boldsymbol{s}_t, \boldsymbol{a}(\boldsymbol{s}_t)) = c_1(\boldsymbol{s}_t, \boldsymbol{\phi}(\boldsymbol{s}_t)) + c_2(\boldsymbol{s}_t, \boldsymbol{\psi}(\boldsymbol{s}_t)), \;\; \boldsymbol{s}_t \in \mathcal{S}, \tag{15}$$

where

$$\begin{aligned} c_1(\boldsymbol{s}_t, \boldsymbol{\phi}(\boldsymbol{s}_t)) = &\sum_{(m,n)\in\mathcal{L}} \mu^{\mathrm{switch}}_{mn}\Big|\alpha_{mn}(\boldsymbol{s}_t) - \alpha_{mn}(\boldsymbol{s}_{t-1})\Big| \\ &+ \sum_{n\in\mathcal{N}^-}\Big(c_{n0}\, u_n(\boldsymbol{s}_t) + \big(1 - d_n(\boldsymbol{s}_t)\big)\,\omega^{\mathrm{load}}_n P_{D_n}(\boldsymbol{s}_t)\Big), \end{aligned}$$

and

$$c_2(\boldsymbol{s}_t, \boldsymbol{\psi}(\boldsymbol{s}_t)) = \sum_{(n,m)\in\mathcal{L}} \eta^{\mathrm{loss}} r_{nm}\ell_{mn}(\boldsymbol{s}_t) + \sum_{n\in\mathcal{N}^-} c_{n1}P_{G_n}(\boldsymbol{s}_t).$$

Decoupling the tasks of obtaining the action vectors $\boldsymbol{\phi}(\boldsymbol{s}_t)$ and $\boldsymbol{\psi}(\boldsymbol{s}_t)$ enables us to define a modified policy $\widetilde{\boldsymbol{\pi}} = (\widetilde{\boldsymbol{\pi}}(\boldsymbol{s}),\, \boldsymbol{s} \in \mathcal{S})$ that determines only an optimal action vector $\boldsymbol{\phi}(\boldsymbol{s}_t)$ consisting of binary decision variables in state $\boldsymbol{s}_t \in \mathcal{S}$ and obtains the cost $c_1(\boldsymbol{s}_t, \boldsymbol{\phi}(\boldsymbol{s}_t))$. Then, given $\boldsymbol{\phi}(\boldsymbol{s}_t)$, the DNO solves an OPF problem to obtain the action vector $\boldsymbol{\psi}^*(\boldsymbol{s}_t)$ that satisfies power flow constraints (3a)−(11). Let $\Psi_{\boldsymbol{\phi}(\boldsymbol{s}_t)}(\boldsymbol{s}_t)$ denote the feasible action space defined by constraints (3a)−(11) for the given $\boldsymbol{\phi}(\boldsymbol{s}_t)$ in state $\boldsymbol{s}_t$. The DNO solves the following optimization problem to obtain the optimal variable $\boldsymbol{\psi}(\boldsymbol{s}_t)$ for the given vector $\boldsymbol{\phi}(\boldsymbol{s}_t)$ and modified policy $\widetilde{\boldsymbol{\pi}}$:

$$\boldsymbol{\mathcal{P}}^{\mathrm{opf},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}: \quad \underset{\boldsymbol{\psi}(\boldsymbol{s}_t)}{\text{minimize}}\;\; c_2(\boldsymbol{s}_t, \boldsymbol{\psi}(\boldsymbol{s}_t))$$

$$\text{subject to } \boldsymbol{\psi}(\boldsymbol{s}_t) \in \Psi_{\boldsymbol{\phi}(\boldsymbol{s}_t)}(\boldsymbol{s}_t).$$

The modified policy $\widetilde{\boldsymbol{\pi}}$ aims to minimize the expected average cost in (13). Hence, in problem $\boldsymbol{\mathcal{P}}^{\text{MDP}}$, we can replace the original policy $\boldsymbol{\pi}$ with the modified policy $\widetilde{\boldsymbol{\pi}}$ to determine the optimal value of the expected average cost $\rho^{\widetilde{\boldsymbol{\pi}}}$ and relative value function $V^{\widetilde{\boldsymbol{\pi}}}(\boldsymbol{s})$, $\boldsymbol{s} \in \mathcal{S}$. Obtaining the modified policy $\widetilde{\boldsymbol{\pi}}$ that solves problem $\boldsymbol{\mathcal{P}}^{\text{MDP}}$ is still challenging, since the state transition probabilities $\Pr(\boldsymbol{s}' \,|\, \boldsymbol{s}, \boldsymbol{a}(\boldsymbol{s}))$, $\boldsymbol{s}, \boldsymbol{s}' \in \mathcal{S}$ may not be available. In the next section, we develop a DRL-based algorithm based on actor-critic method to gradually update both the relative value function and the modified policy without any knowledge of the state transition probabilities.

III. Algorithm Design

In this section, we propose a DRL algorithm based on actor-critic method that learns and improves the policy through interaction with the dynamic changes of the distribution system to solve problem $\boldsymbol{\mathcal{P}}^{\text{MDP}}$. Our proposed DRL algorithm is a compelling approach to solve problem $\boldsymbol{\mathcal{P}}^{\text{MDP}}$. Unlike standard optimization-based approaches, it can effectively manage uncertainty issues, inter-temporal constraints, and large number of variables, all pertinent to distribution networks. The algorithm gradually learns from the network's dynamics to optimize the use of remote-control switches, backup generators, and interruptible loads to achieve long-term economic efficiency and operational reliability of the distribution network. In the proposed DRL algorithm, there are two DNNs corresponding to the actor and critic, respectively. The actor DNN obtains action $\boldsymbol{\phi}(\boldsymbol{s}_t)$ in state $\boldsymbol{s}_t \in \mathcal{S}$, and the critic DNN assesses these actions by estimating the relative value function, guiding the actor towards better policy decisions. Fig. 1 summarizes the interactions between the actor and critic DNNs during the training phase.

A. Motivation and Overview

In our proposed DRL algorithm, the actor network utilizes a transformer architecture. The promising computational benefits of transformer networks in solving large-scale combinatorial optimization problems (in, e.g., [38], [39]) motivate us to apply such an architecture to model the actor DNN. The top portion of Fig. 1 details the proposed transformer architecture for the actor DNN. The transformer network has an encoder-decoder architecture with attention mechanism, which can obtain the modified policy $\widetilde{\boldsymbol{\pi}}$ that solves problem $\boldsymbol{\mathcal{P}}^{\text{MDP}}$. By adapting the transformer's sequence-to-sequence prediction capabilities, the actor DNN can capture the complexities of the distribution network's topology and operating conditions. For the actor DNN, the input sequence, which encapsulates the current state of the distribution network, is transformed through the attention and feed-forward layers to produce an output sequence that represents the binary control decisions for the switches, generators, and loads. Subsequently, the continuous variables are obtained by solving an OPF problem to ensure adherence to the network's operating constraints. The critic DNN processes the system's state and outputs the relative value function, which evaluates the policy's performance by predicting the expected return. By embedding the transformer architecture within the actor DNN, we enable the DRL algorithm to learn and adapt policies that are both feasible and near-optimal for the DNR problem.

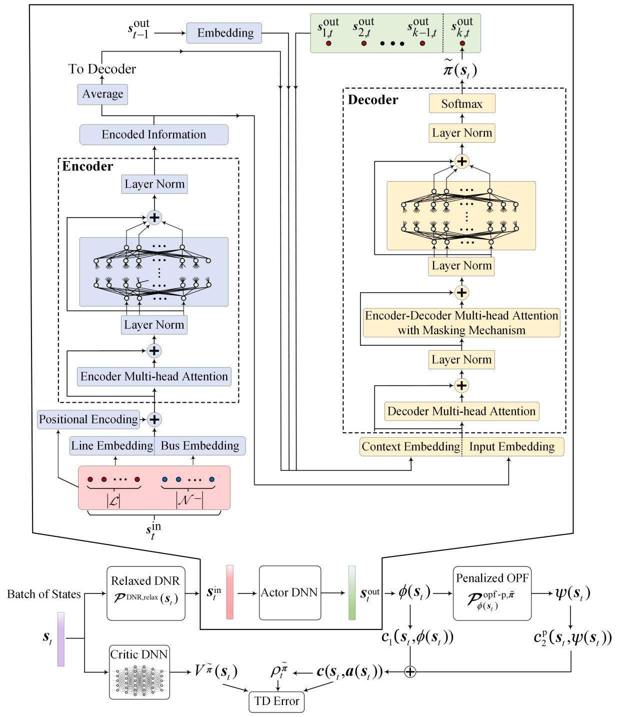


Figure 1. Illustration of the proposed algorithm and the interaction between the actor and critic DNNs. The actor DNN is based on transformer architecture consisting of an encoder and a decoder with multi-head attention mechanism. The critic DNN is based on multi-layer perceptron architecture.

B. Algorithm Training

We denote the network parameter vector of the actor DNN by $\boldsymbol{\theta}$. We use a DNN with MLP architecture and network parameter $\boldsymbol{\vartheta}$ for the critic that receives the system state $\boldsymbol{s}_t \in \mathcal{S}$ as the input and returns the relative value function $V^{\widetilde{\boldsymbol{\pi}}}(\boldsymbol{s}_t, \boldsymbol{\vartheta})$ as the output. Without loss of generality, we can set the training epoch for the actor and critic DNNs to one time slot. To train the actor and critic DNNs for a sufficiently large number of states, we consider a batch $\mathcal{S}_t^{\text{train}} \subseteq \mathcal{S}$ of system states in training epoch t. The DNO observes the system state $\boldsymbol{s}_t \in \mathcal{S}_t^{\text{train}}$ and computes the temporal difference (TD) error $\delta(\boldsymbol{\vartheta}_{t-1}, \boldsymbol{s}_t, \boldsymbol{s}_{t-1})$ for the critic network parameter $\boldsymbol{\vartheta}_{t-1}$, current state $\boldsymbol{s}_t$, and previous state $\boldsymbol{s}_{t-1}$. We have

$$\delta(\boldsymbol{s}_t, \boldsymbol{s}_{t-1}, \boldsymbol{\vartheta}_{t-1}) = c(\boldsymbol{s}_{t-1}, \boldsymbol{a}(\boldsymbol{s}_{t-1})) - \rho_{t-1}^{\widetilde{\boldsymbol{\pi}}} + V^{\widetilde{\boldsymbol{\pi}}}(\boldsymbol{s}_t, \boldsymbol{\vartheta}_{t-1}) - V^{\widetilde{\boldsymbol{\pi}}}(\boldsymbol{s}_{t-1}, \boldsymbol{\vartheta}_{t-1}). \quad (16)$$

We use $\boldsymbol{\delta}(\boldsymbol{\vartheta}_{t-1}) = (\delta(\boldsymbol{s}_t, \boldsymbol{s}_{t-1}, \boldsymbol{\vartheta}_{t-1}),\ \boldsymbol{s}_t \in \mathcal{S}_t^{\text{train}},\ \boldsymbol{s}_{t-1} \in \mathcal{S}_{t-1}^{\text{train}})$ to denote the vector of TD errors. The network parameters for the actor and critic DNNs are updated as follows:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \gamma_t^{\text{a}}\, \boldsymbol{\delta}^{\text{T}}(\boldsymbol{\vartheta}_{t-1})\, \nabla_{\boldsymbol{\theta}} \ln \widetilde{\boldsymbol{\pi}}\big(\boldsymbol{\phi}(\boldsymbol{s}_{t-1}) \,\big|\, \boldsymbol{s}_{t-1}, \boldsymbol{\theta}\big)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{t-1}}, \quad (17a)$$

$$\boldsymbol{\vartheta}_t = \boldsymbol{\vartheta}_{t-1} + \gamma_t^{\text{c}}\, \boldsymbol{\delta}^{\text{T}}(\boldsymbol{\vartheta}_{t-1}) \nabla_{\boldsymbol{\vartheta}} V^{\widetilde{\boldsymbol{\pi}}}(\boldsymbol{s}_{t-1}, \boldsymbol{\vartheta})\Big|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_{t-1}}, \quad (17b)$$

where ∇ is the gradient operator and $\boldsymbol{\delta}^{\mathrm{T}}(\cdot)$ denotes the transpose of vector $\boldsymbol{\delta}(\cdot)$. Parameters γ_t^{a} and γ_t^{c}, respectively, are the step size for the actor and critic updates in time slot t.

C. Input Sequence for the Actor DNN

In general, the input sequence for the actor DNN should provide information about the switches, backup generators, load demands, and the topology of the distribution network. We construct the input sequence $\boldsymbol{s}_t^{\mathrm{in}}$ from the solution to the relaxed version of the DNR problem in state $\boldsymbol{s}_t$. The relaxed DNR problem provides a lower bound for the global optimal solution to the original DNR problem. The optimal solution to the relaxed DNR problem has implicit information about the open/closed state of the switches, on/off state of the backup generators, and connected/disconnected state of the loads in state $\boldsymbol{s}_t$. Hence, the optimal solution to the relaxed DNR problem is a suitable choice for the input sequence $\boldsymbol{s}_t^{\mathrm{in}}$. To formulate the relaxed DNR problem, we relax constraint (1) to allow variables $\alpha_{mn}(\boldsymbol{s}_t)$, $(m,n) \in \mathcal{L}$, $u_n(\boldsymbol{s}_t)$ and $d_n(\boldsymbol{s}_t)$, $n \in \mathcal{N}^-$, to take values within the interval [0, 1]. Thus, we have

$$0 \leq \alpha_{mn}(\boldsymbol{s}_t),\, u_n(\boldsymbol{s}_t),\, d_n(\boldsymbol{s}_t) \leq 1,\; (m,n)\in\mathcal{L},\; n\in\mathcal{N}^-. \quad (18)$$

Furthermore, we relax equality constraint (10) and include the following inequality constraint in the constraint set:

$$\ell_{mn}(\boldsymbol{s}_t)\, z_m(\boldsymbol{s}_t) \geq P_{mn}(\boldsymbol{s}_t)^2 + Q_{mn}(\boldsymbol{s}_t)^2,\; (m,n) \in \mathcal{L}. \quad (19)$$

We construct the feasible action space $\mathcal{A}^{\mathrm{relax}}(\boldsymbol{s}_t)$ from the original action space $\mathcal{A}(\boldsymbol{s}_t)$ by replacing constraints (1) and (10) by constraints (18) and (19), respectively. We solve the following relaxed DNR in system state $\boldsymbol{s}_t \in \mathcal{S}_t^{\mathrm{train}}$:

$$\begin{aligned} \boldsymbol{\mathcal{P}}^{\mathrm{DNR,relax}}(\boldsymbol{s}_t):\quad & \underset{\boldsymbol{a}(\boldsymbol{s}_t)}{\text{minimize}} \;\; c(\boldsymbol{s}_t, \boldsymbol{a}(\boldsymbol{s}_t)) \\ & \text{subject to} \;\; \boldsymbol{a}(\boldsymbol{s}_t) \in \mathcal{A}^{\mathrm{relax}}(\boldsymbol{s}_t). \end{aligned}$$

The relaxed DNR problem $\boldsymbol{\mathcal{P}}^{\mathrm{DNR,relax}}(\boldsymbol{s}_t)$ is an SOCP and can be solved efficiently with polynomial time complexity. Let $\boldsymbol{a}^{\mathrm{relax}}(\boldsymbol{s}_t)$ denote the optimal solution to problem $\boldsymbol{\mathcal{P}}^{\mathrm{DNR,relax}}(\boldsymbol{s}_t)$. From $\boldsymbol{a}^{\mathrm{relax}}(\boldsymbol{s}_t)$, we can extract $\boldsymbol{\phi}^{\mathrm{relax}}(\boldsymbol{s}_t) = ((\alpha_{mn}^{\mathrm{relax}}(\boldsymbol{s}_t),\, (m,n) \in \mathcal{L}),\, (u_n^{\mathrm{relax}}(\boldsymbol{s}_t),\, d_n^{\mathrm{relax}}(\boldsymbol{s}_t),\, n \in \mathcal{N}^-))$ to construct the input sequence $\boldsymbol{s}_t^{\mathrm{in}}$. Thus, as shown in Fig. 1, the input sequence passes through the line and bus embeddings in the encoder that, respectively, convert $(\alpha_{mn}^{\mathrm{relax}}(\boldsymbol{s}_t),\, (m,n) \in \mathcal{L})$ and $(u_n^{\mathrm{relax}}(\boldsymbol{s}_t),\, d_n^{\mathrm{relax}}(\boldsymbol{s}_t),\, n \in \mathcal{N}^-)$ to a high dimensional embedding through a linear projection. The encoder returns the embedded information to the decoder. The decoder uses masking mechanism in the attention layer to account for inter-temporal constraints (e.g., (3a) and (4)). The decoder is executed repeatedly for $|\mathcal{L}| + 2|\mathcal{N}^-|$ times to obtain the output sequence $\boldsymbol{s}_t^{\mathrm{out}}$ consisting of the binary decision variables for the status of the switches, backup diesel generators, and loads.

D. Feasible Power Flow

For the given vector $\boldsymbol{\phi}(\boldsymbol{s}_t)$, the OPF problem $\boldsymbol{\mathcal{P}}^{\mathrm{opf},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$ is solved to determine the action vector $\boldsymbol{\psi}(\boldsymbol{s}_t)$ in state $\boldsymbol{s}_t$. However, problem $\boldsymbol{\mathcal{P}}^{\mathrm{opf},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$ may not have a feasible solution for the given $\boldsymbol{\phi}(\boldsymbol{s}_t)$. We address the infeasibility of problem $\boldsymbol{\mathcal{P}}^{\mathrm{opf},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$ during the training process. We introduce decision variable $\Delta\alpha_{mn}(\boldsymbol{s}_t)$, $\Delta u_n(\boldsymbol{s}_t)$, and $\Delta d_n(\boldsymbol{s}_t)$ for the switch on line $(m,n) \in \mathcal{L}$, and the generator and load connected to bus $n \in \mathcal{N}^-$, respectively. Let vector $\boldsymbol{\psi}^{\mathrm{p}}(\boldsymbol{s}_t) = (\boldsymbol{\psi}(\boldsymbol{s}_t),\, (\Delta\alpha_{mn}(\boldsymbol{s}_t),\, (m,n) \in \mathcal{L}),\, (\Delta u_n(\boldsymbol{s}_t),\, \Delta d_n(\boldsymbol{s}_t),\, n \in \mathcal{N}^-))$ denote the new decision variable. For the given vector $\boldsymbol{\phi}(\boldsymbol{s}_t)$, we construct a new action space $\Psi^{\mathrm{p}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}(\boldsymbol{s}_t)$ in three steps. First, we include the following constraints in the constraint set:

$$0 \leq \alpha_{mn}(\boldsymbol{s}_t) + \Delta\alpha_{mn}(\boldsymbol{s}_t) \leq 1, \qquad (m,n) \in \mathcal{L}, \quad (20a)$$

$$0 \leq u_n(\boldsymbol{s}_t) + \Delta u_n(\boldsymbol{s}_t) \leq 1, \qquad n \in \mathcal{N}^-, \quad (20b)$$

$$0 \leq d_n(\boldsymbol{s}_t) + \Delta d_n(\boldsymbol{s}_t) \leq 1, \qquad n \in \mathcal{N}^-. \quad (20c)$$

Second, in (2)$-$(11), we replace variables $\alpha_{mn}(\boldsymbol{s}_t)$, $u_n(\boldsymbol{s}_t)$, and $d_n(\boldsymbol{s}_t)$ with $\alpha_{mn}(\boldsymbol{s}_t)+\Delta\alpha_{mn}(\boldsymbol{s}_t)$, $u_n(\boldsymbol{s}_t)+\Delta u_n(\boldsymbol{s}_t)$, and $d_n(\boldsymbol{s}_t)+\Delta d_n(\boldsymbol{s}_t)$ for $(m,n) \in \mathcal{L}$, $n \in \mathcal{N}^-$. Third, we replace constraint (10) by constraint (19). We then define a modified objective function $c_2^{\mathrm{p}}(\boldsymbol{s}_t, \boldsymbol{\psi}^{\mathrm{p}}(\boldsymbol{s}_t))$ in state $\boldsymbol{s}_t$ as follows:

$$\begin{aligned} c_2^{\mathrm{p}}(\boldsymbol{s}_t, \boldsymbol{\psi}^{\mathrm{p}}(\boldsymbol{s}_t)) = c_2(\boldsymbol{s}_t, \boldsymbol{\psi}(\boldsymbol{s}_t)) + \kappa\big(\textstyle\sum_{(m,n)\in\mathcal{L}} \left|\Delta\alpha_{mn}(\boldsymbol{s}_t)\right| \\ + \textstyle\sum_{n\in\mathcal{N}^-}\big(\left|\Delta u_n(\boldsymbol{s}_t)\right| + \left|\Delta d_n(\boldsymbol{s}_t)\right|\big)\big), \quad (21) \end{aligned}$$

where κ is a positive weight coefficient. The penalty term in (21) with weighting coefficient κ enables Algorithm 1 (to be presented in the next subsection) to distinguish between feasible and infeasible solutions of the DNR problem during the training process. However, if κ is set to a large number, then Algorithm 1 may converge to a policy with higher expected average cost, because a large penalty for infeasible solutions avoid the actor DNN to explore and obtain a policy with lower expected average cost. Problem $\boldsymbol{\mathcal{P}}^{\mathrm{opf},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$ is transformed into the following optimization problem:

$$\begin{aligned} \boldsymbol{\mathcal{P}}^{\mathrm{opf\text{–}p},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}:\quad & \underset{\boldsymbol{\psi}^{\mathrm{p}}(\boldsymbol{s}_t)}{\text{minimize}} \;\; c_2^{\mathrm{p}}(\boldsymbol{s}_t, \boldsymbol{\psi}^{\mathrm{p}}(\boldsymbol{s}_t)) \\ & \text{subject to} \;\; \boldsymbol{\psi}^{\mathrm{p}}(\boldsymbol{s}_t) \in \Psi^{\mathrm{p}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}(\boldsymbol{s}_t). \end{aligned}$$

Problem $\boldsymbol{\mathcal{P}}^{\mathrm{opf\text{–}p},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$ always has a feasible solution for a given $\boldsymbol{\phi}(\boldsymbol{s}_t)$. We have the following proposition.

Proposition 1: If the original AC OPF problem $\boldsymbol{\mathcal{P}}^{\mathrm{opf},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$ is feasible for $\boldsymbol{\phi}(\boldsymbol{s}_t)$, then by increasing the weight coefficient κ, the global optimal solution of problem $\boldsymbol{\mathcal{P}}^{\mathrm{opf\text{–}p},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$ approaches the global optimal solution of the original problem $\boldsymbol{\mathcal{P}}^{\mathrm{opf},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$.

Proof: Suppose that the original AC OPF problem $\boldsymbol{\mathcal{P}}^{\mathrm{opf},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$ is feasible for $\boldsymbol{\phi}(\boldsymbol{s}_t)$ and let $\boldsymbol{\psi}^*_{\boldsymbol{\phi}(\boldsymbol{s}_t)}(\boldsymbol{s}_t)$ denote its global optimal solution. Define vector $\boldsymbol{\psi}^{\mathrm{p},*}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}(\boldsymbol{s}_t) = (\boldsymbol{\psi}^*_{\boldsymbol{\phi}(\boldsymbol{s}_t)}(\boldsymbol{s}_t),\, (\Delta\alpha_{mn}(\boldsymbol{s}_t),\, (m,n) \in \mathcal{L}),\, (\Delta u_n(\boldsymbol{s}_t),\, \Delta d_n(\boldsymbol{s}_t),\, n \in \mathcal{N}^-))$ for $\Delta\alpha_{mn}(\boldsymbol{s}_t) = 0$, $(m,n) \in \mathcal{L}$, $\Delta u_n(\boldsymbol{s}_t) = 0$, and $\Delta d_n(\boldsymbol{s}_t) = 0$, $n \in \mathcal{N}^-$. Vector $\boldsymbol{\psi}^{\mathrm{p},*}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}(\boldsymbol{s}_t)$ is a feasible solution to the penalized OPF problem $\boldsymbol{\mathcal{P}}^{\mathrm{opf\text{–}p},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$. Given that κ exceeds the norm of the gradient at $\boldsymbol{\psi}(\boldsymbol{s}_t) = \boldsymbol{\psi}^*_{\boldsymbol{\phi}(\boldsymbol{s}_t)}(\boldsymbol{s}_t)$, i.e., $\kappa > \big|\big|\nabla_{\boldsymbol{\psi}(\boldsymbol{s}_t)} c_2(\boldsymbol{s}_t, \boldsymbol{\psi}(\boldsymbol{s}_t))\big|\big|_{\boldsymbol{\psi}(\boldsymbol{s}_t) = \boldsymbol{\psi}^*_{\boldsymbol{\phi}(\boldsymbol{s}_t)}(\boldsymbol{s}_t)}$, then any divergence from the feasible solution $\boldsymbol{\psi}^*_{\boldsymbol{\phi}(\boldsymbol{s}_t)}(\boldsymbol{s}_t)$, characterized by nonzero deviations in $\Delta\alpha_{mn}(\boldsymbol{s}_t)$ for $(m,n) \in \mathcal{L}$, $\Delta u_n(\boldsymbol{s}_t)$, or $\Delta d_n(\boldsymbol{s}_t)$ for $n \in \mathcal{N}^-$, would result in a larger objective

Algorithm 1: DRL-based DNR Algorithm.

```
1  Set $t := 1$, $\varepsilon := 10^{-6}$.
2  Initialize neural network parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\vartheta}_1$ randomly.
3  Select the batch $\mathcal{S}_1^{\text{train}} \subseteq \mathcal{S}$ of initial system states randomly.
4  repeat
5      Observe the system state $\boldsymbol{s}_t \in \mathcal{S}_t^{\text{train}}$.
6      if $t \neq 1$ then
7          Determine the TD error according to (16).
8          Obtain the updated $\boldsymbol{\theta}_t$ according to (17a).
9          Obtain the updated $\boldsymbol{\vartheta}_t$ according to (17b).
10     end
11     Compute input sequence $\boldsymbol{s}_t^{\text{in}}$ corresponding to state $\boldsymbol{s}_t \in \mathcal{S}_t^{\text{train}}$
         for the actor DNN by solving problem $\boldsymbol{\mathcal{P}}^{\text{DNR,relax}}(\boldsymbol{s}_t)$.
12     Set $\boldsymbol{\phi}(\boldsymbol{s}_t)$, $\boldsymbol{s}_t \in \mathcal{S}_t^{\text{train}}$ to the output sequence $\boldsymbol{s}_t^{\text{out}}$, which is
         obtained by forward propagation in the actor DNN.
13     Obtain $\boldsymbol{\psi}^{\text{p}}(\boldsymbol{s}_t)$, $\boldsymbol{s}_t \in \mathcal{S}_t^{\text{train}}$ by solving problem $\boldsymbol{\mathcal{P}}^{\text{opf-p},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$.
14     Compute immediate cost $c(\boldsymbol{s}_t, \boldsymbol{a}(\boldsymbol{s}_t))$ for state $\boldsymbol{s}_t \in \mathcal{S}_t^{\text{train}}$.
15     Obtain the updated $\rho_t^{\widetilde{\boldsymbol{\pi}}}$ according to (22).
16     $t := t + 1$.
17 until $||\,\boldsymbol{\delta}(\boldsymbol{\vartheta}_{t-1}) - \boldsymbol{\delta}(\boldsymbol{\vartheta}_{t-2})\,|| \leq \varepsilon,\ t \geq 3$;
```

function value for problem $\boldsymbol{\mathcal{P}}^{\text{opf-p},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$. This completes the proof that $\boldsymbol{\psi}^{\text{p},*}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}(\boldsymbol{s}_t)$ is the global optimal solution of problem $\boldsymbol{\mathcal{P}}^{\text{opf-p},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$. ■

The DNO solves problem $\boldsymbol{\mathcal{P}}^{\text{opf-p},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$. Obtaining the immediate cost $c(\boldsymbol{s}_t, \boldsymbol{a}(\boldsymbol{s}_t)) = c_1(\boldsymbol{s}_t, \boldsymbol{\phi}(\boldsymbol{s}_t)) + c_2^{\text{p}}(\boldsymbol{s}_t, \boldsymbol{\psi}^{\text{p}}(\boldsymbol{s}_t))$, we can compute the average immediate cost $\bar{c}_t = \frac{1}{|\mathcal{S}_t^{\text{train}}|} \sum_{\boldsymbol{s}_t \in \mathcal{S}_t^{\text{train}}} c(\boldsymbol{s}_t, \boldsymbol{a}(\boldsymbol{s}_t))$. Then, the expected average cost is updated as follows:

$$\rho_t^{\widetilde{\boldsymbol{\pi}}} = \rho_{t-1}^{\widetilde{\boldsymbol{\pi}}} + \gamma_t^{\text{e}} \left(\bar{c}_t - \rho_{t-1}^{\widetilde{\boldsymbol{\pi}}} \right), \tag{22}$$

where γ_t^{e} is the step size in time slot t.

E. Algorithm Description

Algorithm 1 describes our proposed DNR algorithm. Lines 1 and 2 correspond to the algorithm initialization. In Line 3, we randomly select the batch $\mathcal{S}_1^{\text{train}} \subseteq \mathcal{S}$ of initial system states in time slot $t = 1$. In Line 5, the DNO observes the system state $\boldsymbol{s}_t \in \mathcal{S}_t^{\text{train}}$ in current time slot t. If time slot $t = 1$, then Lines 7 to 9 for the actor and critic update are not executed. For $t > 1$, in Line 7, the DNO computes the TD error $\delta(\boldsymbol{\vartheta}_{t-1}, \boldsymbol{s}_t, \boldsymbol{s}_{t-1})$ according to (16). In Lines 8 and 9, the network parameters for the actor and critic DNNs are updated according to (17a) and (17b), respectively. In Line 11, problem $\boldsymbol{\mathcal{P}}^{\text{DNR,relax}}(\boldsymbol{s}_t)$ is solved to obtain the input sequence $\boldsymbol{s}_t^{\text{in}}$ for the actor DNN. In Line 12, the output sequence $\boldsymbol{s}_t^{\text{out}}$ is obtained by the forward propagation in the actor DNN. In Line 13, the penalized OPF problem $\boldsymbol{\mathcal{P}}^{\text{opf-p},\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{\phi}(\boldsymbol{s}_t)}$ is solved to obtain $\boldsymbol{\psi}^{\text{p}}(\boldsymbol{s}_t)$, $\boldsymbol{s}_t \in \mathcal{S}_t^{\text{train}}$. In Line 14, the immediate cost $c(\boldsymbol{s}_t, \boldsymbol{a}(\boldsymbol{s}_t))$ for state $\boldsymbol{s}_t \in \mathcal{S}_t^{\text{train}}$ is determined. In Line 15, the expected average cost $\rho_t^{\widetilde{\boldsymbol{\pi}}}$ is updated according to (22). Next time slot begins in Line 16. In Line 17, the stopping criterion for Algorithm 1 is given. In an offline training approach, the infeasibility of a solution does not halt our proposed algorithm, allowing continuous updates of neural network parameters as per Lines 8 and 9 of Algorithm 1. However, in real-time operation of the distribution network, if the trained actor DNN outputs an infeasible action, we can leverage the softmax layer to iterate forward propagation until a feasible action is achieved. The effectiveness of this approach is demonstrated via numerical case studies presented in Section IV-B.

In the design of Algorithm 1, we assume complete information about the network topology and transmission parameters. While this enables well-informed decisions, the DNO may not have complete or up-to-date information. Continuous learning and adaptation are essential for the algorithm's long-term efficacy, accommodating changes in load patterns, renewable generation, and network architecture. Additionally, while simulations provide a controlled environment for validation, real-world deployment requires evaluation of the obtained solution to guarantee robustness and applicability.

IV. Performance Evaluation

In this section, we first evaluate the performance of the proposed DRL-based algorithm in solving the DNR problem for an 119-bus distribution system, which has one substation and 133 branches including 118 sectionalizing switches and 15 tie line switches. Fig. 2 shows one-line diagram of the test system. The data for the test system is sourced from [45]. We assume all lines are switchable to evaluate Algorithm 1 under a worst-case scenario with large action space. The duration of each time slot is set to 15 minutes. This system has approximately 4×10^{15} feasible topological configurations, which demonstrates the complexity of solving the DNR problem to obtain the optimal configuration. We set t^{sw} to be equal to 2 for all switches. We use the per-unit (pu) system for our analysis. The base power of the system is 500 MVA. The limit for the current magnitude of lines is set to 1.05 pu. The lower limit and upper limit on the bus voltages are 0.95 pu and 1.05, respectively. In the objective function of the DNR problem, we set the weighting coefficients $\eta^{\text{loss}} = \$500/\text{MW}$, $\mu_{mn}^{\text{switch}} = \$50$, $(m,n) \in \mathcal{L}$. The weighting coefficient ω_n^{load}, $n \in \mathcal{N}^-$ is uniformly sampled at random from the interval [\$1/kW, \$10/kW] to obtain the cost of load interruption. For the backup diesel generator at bus $n \in \mathcal{N}^-$, we obtain the generation cost coefficients c_{n0} and c_{n1}, $n \in \mathcal{N}^-$ at random from normal distributions with mean values \$1 and \$0.1/kW, and standard deviations \$0.5 and \$0.05/kW, respectively. We set parameter $P_{G_{n,t}}^{\text{max}}$ to 50% of the average active load (over all samples) at that bus. Also, we set parameters r_n^{u} and r_n^{d} to 20% of $P_{G_{n,t}}^{\text{max}}$ for the backup generator at bus $n \in \mathcal{N}^-$. We use the data from [41, Section 5.4] to model the loading capability of the generators for the reactive-power generation.

Obtaining the load demand samples involves two steps. The first step is to scale the active- and reactive-power loads given in [45] and obtain the *daily average load profile* for each bus. As shown in Fig. 3, we consider two different scaling factors corresponding to two types (i.e., type 1 and type 2) of loads. Type 1 captures residential load demands with peak load between 6 pm and 11 pm, and type 2 models commercial load demands with peak load between 11 am and 6 pm. We obtain the average load profiles by randomly scaling the load demands at 75% of the buses with type 1 scaling factor and

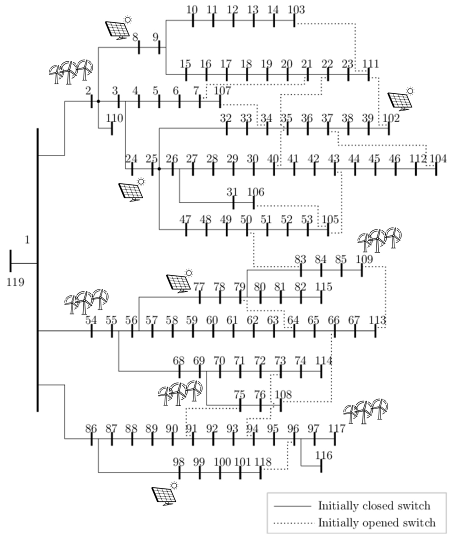


Figure 2. 119-bus test feeder with 118 sectionalizing switches, 15 tie line switches, 5 buses (i.e., buses 2, 54, 68, 102, and 117) with wind generators, and 5 buses (i.e., buses 8, 24, 77, 98, and 102) with PV panel.

at 25% of the buses with type 2. The second step is to use the average load profile at each bus to randomly sample the data for the active load per time slot. That is, we scale the historical load demand data from Ontario, Canada power grid database [46], from January 1, 2021 to March 31, 2022, such that the average value at each time slot is equal to the average load for that bus. Also, we consider a fixed power factor to obtain the samples for reactive load at each bus.

We assume that five buses are connected to PV panels, each with nominal capacity of 300 kWh and five buses are connected to wind turbines, each with nominal capacity of 1000 kWh. To obtain the samples for the output power of the renewable generators, we scale down the available historical data from Ontario, Canada power grid database [46], such that the maximum generation level becomes equal to the aforementioned nominal capacity. To obtain the random samples for maximum active-power supply at substation bus (i.e., bus 119), $P_{G_{n,t}}^{\text{sub,max}}$, we consider two scenarios: (i) the value of $P_{G_{n,t}}^{\text{sub,max}}$ is sufficiently large to meet the total load in all time slots, and (ii) the value of $P_{G_{n,t}}^{\text{sub, max}}$ is lower than the total load in a particular time slot.

For the critic DNN, we consider a multi-layer perceptron consisting of three hidden layers with 10 nodes and leaky rectified linear unit (ReLU) activation function. The actor DNN has a transformer architecture, where the attention layers have 8 heads. The dimension of the bus and line embeddings is set to 128. The dimension of the feed forward layers in the encoder and decoder is set to 512. We perform simulations using Python/PyTorch and Python/CVXPY with MOSEK solver [47] on the Digital Research Alliance of Canada platform [48] with 16 CPUs and 4 GPUs.

A. Algorithm Convergence

We train Algorithm 1 for 7000 iterations. For Algorithm 1 to converge, the update of the actor operates on a slower timescale than the critic, to ensure that the critic has sufficient time to evaluate the current policy. That is, for some positive constant d, we have $\sum_{t=1}^{\infty}(\gamma_t^{\text{a}}/\gamma_t^{\text{c}})^d < \infty$ [49]. Hence, we set the step sizes in Algorithm 1 to $\gamma_t^{\text{c}} = 5/t^{0.5}$, $\gamma_t^{\text{a}} = 10/t$, and $\gamma_t^{\text{e}} = 50/t^{0.5}$. For the training process in Algorithm 1, we consider the batch of $|\mathcal{S}_1^{\text{train}}| = 100$ initial system states randomly. The TD error gradually approaches zero in about 5000 iterations, as shown in Fig. 4(a), which indicates the convergence of Algorithm 1 in the training phase. Meanwhile, Fig. 4(b) shows that the expected average cost converges to \$1581 per time slot.

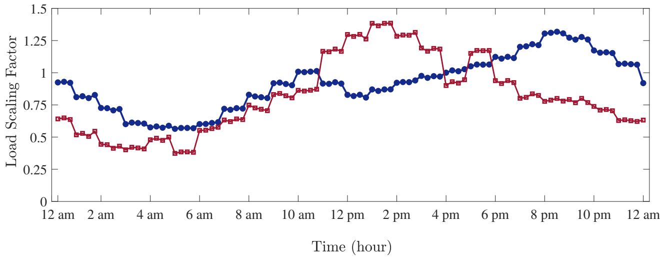
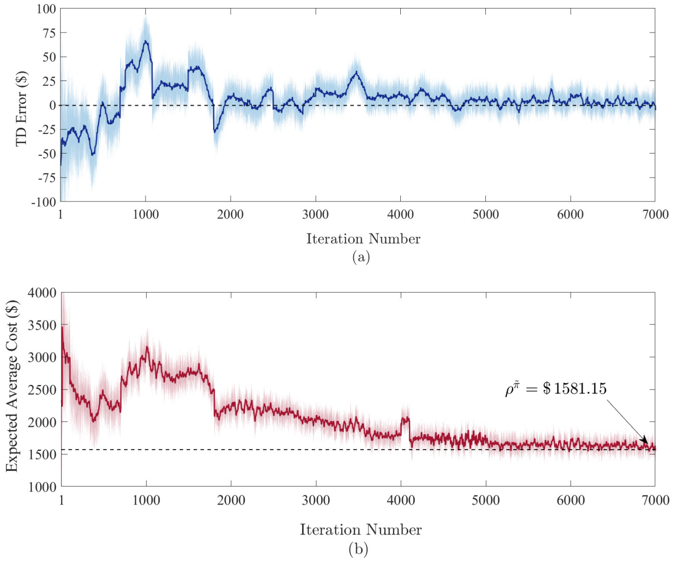


Figure 3. Load scaling factor during one day for type 1 (blue) and type 2 (red) load demands.



(a)

(b)

Figure 4. (a) TD error and (b) expected average cost per time slot versus iteration number for a 119-bus test system.

B. Solution Feasibility and Optimality

We compare the convergence of the expected average cost with $\kappa = 10^2$ and $\kappa = 10^6$ in Fig. 5(a). With $\kappa = 10^2$, the actor DNN obtains a policy with lower expected average cost. With $\kappa = 10^6$, a larger penalty is incurred for infeasible solutions, which enforce the actor DNN not to explore the action space efficiently and obtain a policy with lower expected average cost. In addition to properly set the weighting coefficient κ, we can run forward propagation for multiple rounds in the test process and select the feasible solution with the smallest objective value. In particular, the softmax layer in the output of the actor DNN provides a probability distribution for the possible actions in output sequence. Hence, by performing multiple rounds of forward propagation, the actor DNN can result in multiple choices as the final solution. Then, we

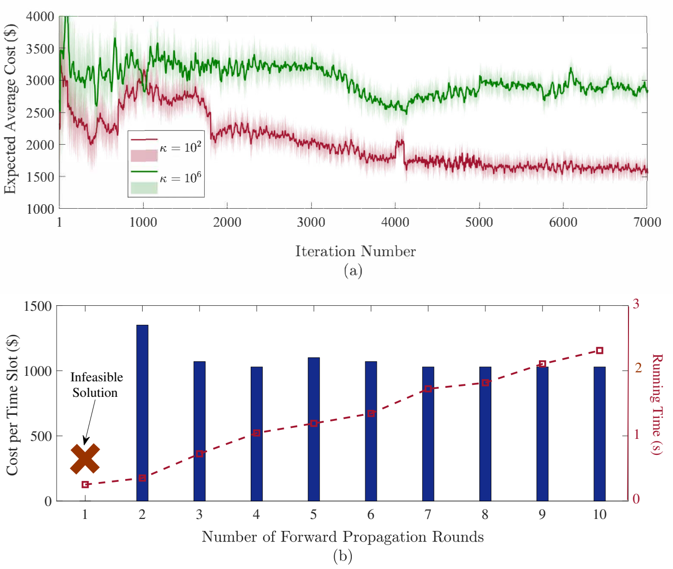


Figure 5. (a) Convergence of the expected average cost with weighting coefficients $\kappa = 10^2$ and $\kappa = 10^4$; and (b) Immediate cost per time slot and algorithm running time versus the number of forward propagation rounds in the actor DNN.

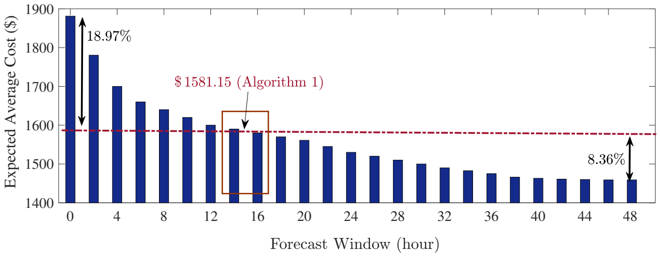


Figure 6. Expected average cost for Algorithm 1 (dashed line) and MISOCP (bar graph) with forecast window between zero and 48 hours.

can select the one with the smallest objective value, or repeat forward propagation until a feasible solution is obtained. To justify this approach, in Fig. 5(b), we show an example for the cost at 6 am versus the number of rounds that we run forward propagation in the actor DNN to select the solution with the lowest objective value. Results demonstrate that if we run forward propagation for one round, then it is possible to obtain an infeasible solution. However, by increasing the number of forward propagation rounds, we can obtain a feasible and near-optimal solution with a smaller objective value, while keeping the running time to be less than 3 seconds.

Next we study the optimality of the solution obtained by Algorithm 1. For comparison, we consider the objective value of the DNR problem $\mathcal{P}^{\text{DNR}}$, when the DNO can forecast the renewable generation and load demand during a pre-specified time window. With complete information about the renewable generation and load demand, problem $\mathcal{P}^{\text{DNR}}$ can be transformed into a deterministic mixed-integer SOCP (MISOCP), which can be solved using MOSEK solver. In Fig. 6, we compare the expected average cost for Algorithm 1 and the deterministic MISOCP with forecast window from zero to 48 hours. When the forecast window is zero, then the DNO solves $\mathcal{P}^{\text{DNR}}$ at the beginning of each time slot $t \in \mathcal{T}$. When the forecast window is 48 hours, then the DNO solves $\mathcal{P}^{\text{DNR}}$ every two days with complete information about the renewable generation and load demand for the next 48 hours. On the other hand, Algorithm 1 is based on DRL that can gradually learn a near-optimal policy to deal with the inter-temporal constraints and stochastic processes behind the uncertain renewable generation and load demand. The results in Fig. 6 show that the objective value obtained by Algorithm 1 is 18.97% lower than the objective value of the deterministic DNR with the forecast window of zero. The gap between the objective values obtained by Algorithm 1 and the objective value of the deterministic DNR with the forecast window of two days is only 8.36%. Moreover, the objective value obtained by Algorithm 1 is approximately equal to the objective value of the deterministic DNR with the forecast window of 16 hours. In other words, with Algorithm 1, the DNR problem can be solved as if we have complete information about the renewable generation and load demand for 16 hours. This demonstrates that Algorithm 1 can effectively address the uncertainty issues and obtain a near-optimal solution to the DNR problem with generation and load uncertainty.

C. Comparison with Existing DNR Algorithms

Next, we compare the performance of Algorithm 1 with the batch-constrained DRL algorithm from [32]. Moreover, we evaluate the effectiveness of our penalized OPF method, outlined in Section III-D, in training the actor and critic DNNs. This approach is compared with the alternative strategies in [35] and [36], which apply penalties to all operational constraint violations. For fair comparison, we apply the algorithm in [32] to solve problem $\mathcal{P}^{\text{MDP}}$ with expected average cost in (13) and relative value function in (14). The algorithm in [32] uses an off-policy approach to learn a control policy from a given set of historical operational data (i.e., configurations and power flow for the underlying distribution network). The historical operational data contain relevant information about the state, action, and cost in the MDP for the DNR problem. The DNO can learn a batch-constrained policy, which is limited to the state-action pairs contained in the historical data. Similar to [32], we assume at most one pair of switch status change per time slot and obtain the historical configuration and power flow data by simulating 10^4 different scenarios.

Fig. 7(a) shows the total load demand and total supplied energy over 10 consecutive days. The total load is greater than the total supplied energy in days 3, 4, 9, and 10. Hence, during some time slots, a number of loads are interrupted and some backup diesel generators are operated to maintain the generation-load balance. In Fig. 7(b), we compare the expected average cost from day one to day ten for the DNR with complete information, Algorithm 1 with DRL, algorithm in [32] with batch-constrained DRL, greedy algorithm, and DNR with penalizing all operation constraints as in [35] and [36]. The DNR with complete information provides a lower bound for the expected average cost, because we solve a deterministic MISOCP with complete information about load demands and renewable sources for ten days. In the greedy method, we solve a deterministic MISOCP at every time slot with limited information about load demand and renewable generation in that time slot. Fig. 7(b) shows that the gap between the expected average cost with our proposed algorithm and the DNR with complete information is 4.7% on average (varies between 2% in days 1 and 6, and 12% in days 3 and 10). The gap is larger for days with greater load demand than supplied energy, since it is more difficult to obtain a near-optimal solution for the network configuration, set of interrupted loads, and set of operating backup generators under uncertainty about load demands and renewable generators. Moreover, Fig. 7(b) shows that the gap between the expected average cost with our proposed algorithm and that in [32] is at least 11% (for day 10). The relatively poor performance of the algorithm in [32] in solving the DNR problem is due to the off-policy approach with historical operational data and the limit of at most one pair of switching action per time slot. However, in Algorithm 1, the proposed policy DNN with transformer architecture can address multiple switching actions to obtain a near-optimal binary solution for the status of the switches.

In our final comparison, we assess the efficacy of using the proposed penalized OPF method detailed in Section III-D to train the actor and critic DNNs against an alternative strategy that penalizes all operational constraint violations, as employed in the proposed safe DRL algorithm in [35] and the proposed deep learning algorithm in [36]. This alternative approach involves incorporating violations of operational constraints (4)−(11) directly into the objective function as penalties. While both techniques facilitate the gradual learning of feasible solutions by the DNNs, our findings, as illustrated in Fig. 7(b), reveal a significant advantage in using the penalized OPF approach, which achieves a 24.5% lower objective value on average. The poorer performance of the method that penalizes all operating constraints is due to its overly conservative exploration strategy that any violation incurs substantial penalties, making the learning process sensitive to constraint violations. In contrast, our proposed algorithm, by focusing penalties solely on infeasible binary solutions within the OPF framework, effectively guides the neural network in learning to select feasible actions for switch operations, load management, and the activation of backup diesel generators.

D. Demonstrating Algorithm Scalability

To study the scalability of Algorithm 1, we construct test systems with 595, 1190, and 3570 buses by connecting 5, 10, and 30 119-bus feeders, respectively, via branches between two random buses with tie line switch, resistance of 0.01 pu, and reactance of 0.001 pu. This scenario helps to demonstrate the performance of Algorithm 1 in large test systems with multiple substation buses. Fig. 8 shows that, by using Algorithm 1, the expected average cost per time slot converges to the suboptimal solution in 8200 iterations, 15000 iterations, and 35000 iterations in test systems with 595, 1190, and 3570 buses, respectively. When compared with our original case study (i.e., 119-bus test system), the number of iterations for convergence is higher in these large test systems, because the policy DNN needs to adjust the policy according to the time-varying load demand and renewable generation, and the operation constraints in a larger systems. Despite a higher number of iterations for convergence, Algorithm 1 can still be used for large test systems, since it can be executed in an offline manner given historical data. With appropriate training data, Algorithm 1 can learn the dynamics of different distribution networks varying in size or having different operational constraints. To enhance the adaptability of the proposed algorithm, one can use transfer learning techniques [50], where the knowledge from a small system can serve as a foundation to train the model for a large system.

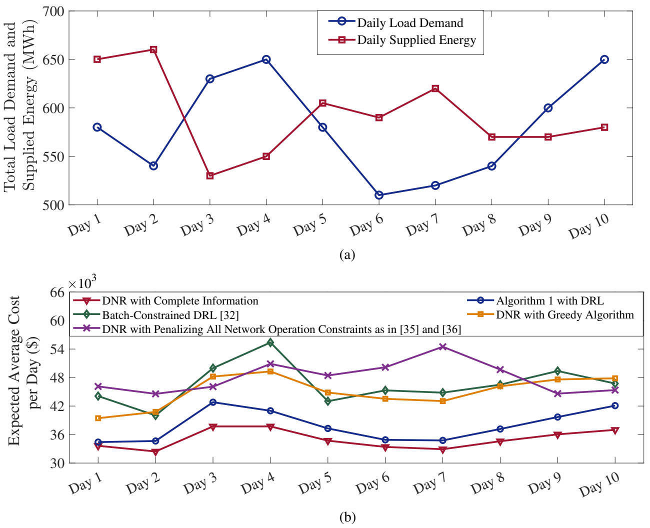


Figure 7. (a) Total daily load demand and total supplied energy; and (b) expected average cost per day for the DNR with complete information, Algorithm 1, algorithm in [32] with batch-constrained DRL, greedy method, and DNR with penalizing all operating constraints as in [35] and [36].

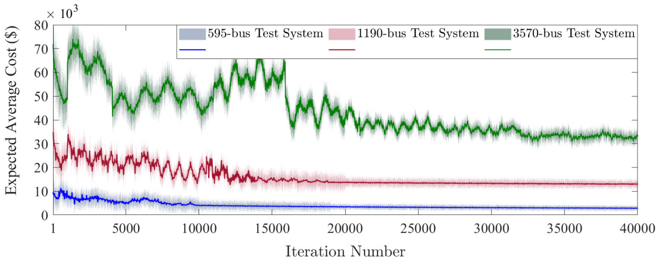


Figure 8. Expected average cost for the test systems with 595, 1190, and 3570 buses.

V. Conclusion

In this paper, we studied DNR problem under uncertainty in the load demand and renewable generation. We accounted for network topology, backup generator operation, emergency load interruption, and distribution network constraints. To address the uncertainty issues, we developed a DRL algorithm with actor-critic method. To obtain binary decision variables, we decoupled the tasks of (i) obtaining the action vector for the status of the switches, backup generators, and loads and (ii) obtaining the action vector for a feasible power flow in the distribution network. We designed a transformer-based architecture for the policy DNN. To address the non-convexity of power flow constraints, we transformed the problem of updating neural network parameters into a sequence of SOCPs. Via numerical simulations in an 119-bus test system, we showed that the fast convergence of the proposed DRL algorithm to a near-optimal solution with an average gap of 4.7% from the optimal solution of the deterministic DNR with complete information about future load demand and renewable generation. Moreover, the expected average cost with our proposed algorithm is at least 11% lower than the expected average cost with an existing DNR algorithm in the literature. Our analysis demonstrated the superiority of the proposed penalized OPF approach with 24.5% lower objective value when compared to the alternative approaches the literature, which impose penalties on all operational constraint violations. Furthermore, the proposed algorithm scales well to solve the DNR problem in large test systems with 595, 1190, and 3570 buses. For future work, we will consider the incomplete information about distribution network architecture, both in formulating the DNR problem and in developing a learning algorithm.

References

[1] H. L. Willis, *Power Distribution Planning Reference Book*, 2nd ed. FL: CRC Press, 2004.

[2] Siemens. [Online]. Available: https://www.siemens.com/global/en/products/automation/industrial-communication/smart-grid-electric-power-utilities/distribution-automation

[3] S&c electric company. [Online]. Available: https://www.sandc.com/en/products–services/products/scada-mate-switching-systems

[4] Schneider electric. [Online]. Available: https://www.se.com/ca/en/product-subcategory/1950-remote-control-and-monitoring

[5] Hubbell power systems. [Online]. Available: https://www.hubbell.com/hubbellpowersystems/en/solutions/advanced-distribution-automation

[6] Gw electric. [Online]. Available: https://www.gwelectric.com/products/distribution-reclosers-and-overhead-switches

[7] Switched source. [Online]. Available: https://www.switchedsource.com/tie-controller

[8] Allied power and control. [Online]. Available: https://alliedpowerandcontrol.com/product/remote-control-switches

[9] Caterpillar. [Online]. Available: https://www.cat.com/en_US/products/new/power-systems/electric-power/switchgear-and-paralleling-controls

[10] H. Sekhavatmanesh and R. Cherkaoui, "A multi-step reconfiguration model for active distribution network restoration integrating DG start-up sequences," *IEEE Trans. on Sustainable Energy*, vol. 11, no. 4, pp. 2879–2888, Oct. 2020.

[11] Q. Peng, Y. Tang, and S. H. Low, "Feeder reconfiguration in distribution networks based on convex relaxation of OPF," *IEEE Trans. on Power Systems*, vol. 30, no. 4, pp. 1793–1804, Jul. 2015.

[12] J. A. Taylor and F. S. Hover, "Convex models of distribution system reconfiguration," *IEEE Trans. on Power Systems*, vol. 27, no. 3, pp. 1407–1413, Aug. 2012.

[13] Y. Liu, J. Li, and L. Wu, "Coordinated optimal network reconfiguration and voltage regulator/DER control for unbalanced distribution systems," *IEEE Trans. on Smart Grid*, vol. 10, no. 3, pp. 2912–2922, May 2019.

[14] C. Wang, P. Ju, S. Lei, Z. Wang, F. Wu, and Y. Hou, "Markov decision process-based resilience enhancement for distribution systems: An approximate dynamic programming approach," *IEEE Trans. on Smart Grid*, vol. 11, no. 3, pp. 2498–2510, May 2020.

[15] A. Akrami, M. Doostizadeh, and F. Aminifar, "Optimal reconfiguration of distribution network using μPMU measurements: A data-driven stochastic robust optimization," *IEEE Trans. on Smart Grid*, vol. 11, no. 1, pp. 420–428, Jan. 2020.

[16] A. Azizivahed, A. Arefi, S. Ghavidel, M. Shafie-khah, L. Li, J. Zhang, and J. P. S. Catalão, "Energy management strategy in dynamic distribution network reconfiguration considering renewable energy resources and storage," *IEEE Trans. on Sustainable Energy*, vol. 11, no. 2, pp. 662–673, Apr. 2020.

[17] W. Zheng, W. Huang, D. J. Hill, and Y. Hou, "An adaptive distributionally robust model for three-phase distribution network reconfiguration," *IEEE Trans. on Smart Grid*, vol. 12, no. 2, pp. 1224–1237, Mar. 2021.

[18] C. Wang, K. Pang, M. Shahidehpour, F. Wen, and S. Duan, "Two-stage robust design of resilient active distribution networks considering random tie line outages and outage propagation," *IEEE Trans. on Smart Grid*, vol. 14, no. 4, pp. 2630–2644, Jul. 2023.

[19] M. R. Dorostkar-Ghamsari, M. Fotuhi-Firuzabad, M. Lehtonen, and A. Safdarian, "Value of distribution network reconfiguration in presence of renewable energy resources," *IEEE Trans. on Power Systems*, vol. 31, no. 3, pp. 1879–1888, May 2016.

[20] X. Cao, J. Wang, J. Wang, and B. Zeng, "A risk-averse conic model for networked microgrids planning with reconfiguration and reorganizations," *IEEE Trans. on Smart Grid*, vol. 11, no. 1, pp. 696–709, Jan. 2020.

[21] M. Rahmani-Andebili and M. Fotuhi-Firuzabad, "An adaptive approach for PEVs charging management and reconfiguration of electrical distribution system penetrated by renewables," *IEEE Trans. on Industrial Informatics*, vol. 14, no. 5, pp. 2001–2010, May 2018.

[22] S. Huang, Q. Wu, L. Cheng, and Z. Liu, "Optimal reconfiguration-based dynamic tariff for congestion management and line loss reduction in distribution networks," *IEEE Trans. on Smart Grid*, vol. 7, no. 3, pp. 1295–1303, May 2016.

[23] C. Lei, S. Bu, J. Zhong, Q. Chen, and Q. Wang, "Distribution network reconfiguration: A disjunctive convex hull approach," *IEEE Trans. on Power Systems*, vol. 38, no. 6, pp. 5926–5929, Nov. 2023.

[24] 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.

[25] H. Wu, P. Dong, and M. Liu, "Distribution network reconfiguration for loss reduction and voltage stability with random fuzzy uncertainties of renewable energy generation and load," *IEEE Trans. on Industrial Informatics*, vol. 16, no. 9, pp. 5655–5666, Sept. 2020.

[26] A. Kavousi-Fard, T. Niknam, and M. Fotuhi-Firuzabad, "A novel stochastic framework based on cloud theory and θ-modified bat algorithm to solve the distribution feeder reconfiguration," *IEEE Trans. on Smart Grid*, vol. 7, no. 2, pp. 740–750, Mar. 2016.

[27] A. Asrari, S. Lotfifard, and M. Ansari, "Reconfiguration of smart distribution systems with time varying loads using parallel computing," *IEEE Trans. on Smart Grid*, vol. 7, no. 6, pp. 2713–2723, Nov. 2016.

[28] S. Razavi, H.-R. Momeni, M.-R. Haghifam, and S. Bolouki, "Multi-objective optimization of distribution networks via daily reconfiguration," *IEEE Trans. on Power Delivery*, vol. 37, no. 2, pp. 775–785, Apr. 2022.

[29] P. Harsh and D. Das, "A simple and fast heuristic approach for the reconfiguration of radial distribution networks," *IEEE Trans. on Power Systems*, vol. 38, no. 3, pp. 2939–2942, May 2023.

[30] Y. Li, G. Hao, Y. Liu, Y. Yu, Z. Ni, and Y. Zhao, "Many-objective distribution network reconfiguration via deep reinforcement learning assisted optimization algorithm," *IEEE Trans. on Power Delivery*, vol. 37, no. 3, pp. 2230–2244, Jun. 2022.

[31] W. Huang, W. Zheng, and D. J. Hill, "Distribution network reconfiguration for short-term voltage stability enhancement: An efficient deep learning approach," *IEEE Trans. on Smart Grid*, vol. 12, no. 6, pp. 5385–5395, Nov. 2021.

[32] Y. Gao, W. Wang, J. Shi, and N. Yu, "Batch-constrained reinforcement learning for dynamic distribution network reconfiguration," *IEEE Trans. on Smart Grid*, vol. 11, no. 6, pp. 5357–5369, Nov. 2020.

[33] B. Wang, H. Zhu, H. Xu, Y. Bao, and H. Di, "Distribution network reconfiguration based on NoisyNet deep Q-learning network," *IEEE Access*, vol. 9, pp. 90 358–90 365, Jun. 2021.

[34] Y. Zhang, F. Qiu, T. Hong, Z. Wang, and F. Li, "Hybrid imitation learning for real-time service restoration in resilient distribution systems," *IEEE Trans. on Industrial Informatics*, vol. 18, no. 3, pp. 2089–2099, Mar. 2022.

[35] H. Li and H. He, "Learning to operate distribution networks with safe deep reinforcement learning," *IEEE Trans. on Smart Grid*, vol. 13, no. 3, pp. 1860–1872, May 2022.

[36] R. Wang, X. Bi, and S. Bu, "Real-time coordination of dynamic network reconfiguration and Volt-VAR control in active distribution network: A graph-aware deep reinforcement learning approach," accepted for publication in *IEEE Trans. on Smart Grid*, Oct. 2023.

[37] S. Bahrami, Y. Chen, and V.W.S. Wong, "A neural combinatorial optimization algorithm for unit commitment in AC power systems," in *Proc. of IEEE SmartGridComm*, Singapore, Oct. 2022.

[38] W. Kool, H. van Hoof, and M. Welling, "Attention, learn to solve routing problems!" in *Proc. of Int'l. Conf. on Learning Representations (ICLR)*, New Orleans, LA, May. 2019.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. of Conf. on Neural Information Processing Systems (NIPS)*, Long Beach, CA, Dec. 2017.

[40] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA: MIT Press, 2018.

[41] P. Kundur, N. Balu, and M. Lauby, *Power System Stability and Control*, 1st ed. NY: McGraw-Hill Science, 1994.

[42] J. D. Glover, M. S. Sarma, and T. Overbye, *Power System Analysis and Design*, 5th ed. CT: Cengage Learning, 2011.

[43] R. Miller and J. Malinowski, *Power System Operation*, 3rd ed. NY: McGraw-Hill Education, 1994.

[44] R. Billinton and R. N. Allen, *Reliability Evaluation of Power Systems*, 2nd ed. MA: Kluwer Boston Incorporated, 1996.

[45] D. Zhang, Z. Fu, and L. Zhang, “An improved TS algorithm for loss-minimum reconfiguration in large-scale distribution systems,” *Electric Power Systems Research*, vol. 77, no. 5, pp. 685–694, Apr. 2007.

[46] Independent Electricty System Operator (IESO). [Online]. Available: www.ieso.ca.

[47] MOSEK. [Online]. Available: www.mosek.com

[48] The Digital Research Alliance of Canada. [Online]. Available: www.alliancecan.ca

[49] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, “A survey of actor-critic reinforcement learning: Standard and natural policy gradients,” *IEEE Trans. on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, vol. 42, no. 6, pp. 1291–1307, Nov. 2012.

[50] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, “A comprehensive survey on transfer learning,” *Proc. of the IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.