Graph-based Safe Reinforcement Learning for Dynamic Optimal Power Flow with Hybrid Action Space Considering Time-varying Network Topologies

Xihai Zhang, *Student Member*, *IEEE*, Shaoyun Ge, Yue Zhou, *Member*, *IEEE*, Hong Liu, *Member*, *IEEE*, Shida Zhang, and Changxu Jiang

***Abstract*—The proliferation of distributed energy resources and time-varying network topologies in active distribution networks presents unprecedented challenges for network operators. While reinforcement learning (RL) has shown promise in addressing network-constrained energy scheduling, it faces difficulties in managing the complexities of dynamic topologies and discrete-continuous hybrid action spaces. To address these challenges, a graph-based safe RL approach is proposed to learn dynamic optimal power flow under time-varying network topologies. This proposed approach leverages graph convolution operators to handle network topology changes, while safe RL with parameterized action ensures policy development. Specifically, the graph convolution operator abstracts key characteristics of the network topology, enabling effective power flow management in non-stationary environments. Besides that, a parameterized action constrained Markov decision process is employed to handle the hybrid action space and ensure compliance with physical network constraints, thereby accelerating the deployment of safe policy for hybrid action spaces. Numerical results demonstrate that the proposed approach efficiently navigates the discrete-continuous decision space while accounting for the constraints imposed by the dynamic nature of power flow in time-varying network topologies.**

***Index Terms*—Active distribution network, distributed energy resource, reinforcement learning, graph convolution operator, network topology, hybrid action space, optimal power flow.**

Manuscript received: November 5, 2024; revised: February 11, 2025; accepted: March 27, 2025. Date of CrossCheck: March 27, 2025. Date of online publication: May 2, 2025.

This work was supported by the Tianjin Science and Technology Program (No. 22JCZDJC00820).

X. Zhang, S. Ge, and H. Liu (corresponding author) are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China. (e-mail: xihaizhang@tju.edu.cn; syge@tju.edu.cn; liuhong@tju.edu.cn).

Y. Zhou is with the Department of Electrical and Electronic Engineering, School of Engineering, Cardiff University, Cardiff, CF24 3AA, UK (e-mail: ZhouY68@cardiff.ac.uk).

S. Zhang is with the School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China (e-mail: zhangshida@zzu.edu.cn).

C. Jiang is with the School of Electrical Engineering and Automation, Fuzhou University, Fuzhou 310108, China (e-mail: cxjiang@fzu.edu.cn).

DOI: 10.35833/MPCE.2024.001198

I. Introduction

OPTIMAL power flow (OPF) is central to power system operation and is viewed as a complex economic, electrical, and computational problem [1]. The ubiquitous distributed energy resources (DERs) within active distribution networks (ADNs), such as photovoltaics (PVs) and energy storage systems (ESSs), emphasize the importance of managing their dynamic nature and uncertainties [2]. The emergence of smart meters and the unprecedentedly large volumes of data have triggered a shift in ADN control from reliance on local control loops to grid-state responsiveness [3].

Viewed from the perspective of ADN operation, the dynamic optimal power flow (DOPF) involves the strategic dispatching of available resources to minimize operating costs and network losses simultaneously across varying time horizons [4]. Distinguished from the static OPF paradigm, DOPF emphasizes the integration of time-couple technologies, wherein decision-making spans multiple time horizons to effectively manage flexible resources such as ESSs. Furthermore, the trend in accommodating the high penetrations of renewable energy introduces significant stochasticity and fluctuations. Consequently, there is a critical need to tackle the inherent uncertainty associated with formulating DOPF to ensure the effectiveness and safety of ADN operations.

Prior works on modeling DOPF with high penetrations of renewable energy have predominantly focused on uncertain-aware mathematical programming, deep learning (DL), and reinforcement learning (RL) approaches. For characterizing renewable energy generation uncertainties, the uncertain-aware mathematical programming approaches concentrate on robust optimization (RO), stochastic optimization (SO), and model predictive control (MPC). The robust OPF methodologies are investigated via leveraging convex hull [5], two-stage adaptive RO [6], and scenario-based RO [7]. Despite RO has the ability to handle uncertainty, it tends to yield conservative solutions as it prioritizes worst-case scenarios within the uncertainty set. The SO-based approaches aim to characterize uncertainties via probabilistic-based or scenario-based methods. The probabilistic-based methods such as chance-constrained [8], conditional value at risk [9], and robust SO [10] introduce a hyperparameter to quantify and manage their tail risks. However, these hyperparameters lack interpretability, making it challenging to understand their risk implications for risk management. The scenario-based methods, like sample average approximation (SAA) [11], aim to approximate uncertainties by generating discrete scenarios for representing potential realizations of uncertain parameters. While these methods are straightforward for estimating the expected value of objectives, they face the curse of dimensionality and low sample efficiency in large-scale optimization problems. The MPC-based framework is presented in [12], [13] to optimize power flow and coordinate demand response jointly within interconnected cluster of ADNs via its predictive and self-correcting capabilities. Nonetheless, MPC necessitates solving optimization problems online, which demands substantial computational resources and relies heavily on accurate system dynamic models. The measurement noise and estimation error could adversely affect long-term performance. It is worth noting that the aforementioned uncertain-aware mathematical programming approaches rely on linearized or second-order cone relaxation to approximate the non-convex nature of power flow. However, these approaches neglect power loss components, resulting in an incomplete representation of ADNs with high ratio of resistance to reactance [14]. Besides that, the original power flow formulation, particularly involving integer variables, is computationally intensive due to its NP-hard nature. Consequently, finding a trade-off solution between computational efficiency and precision in the governing equations of power flow remains a significant challenge.

The DL-based DOPF approach aims to predict the solution of alternating current (AC) OPF problem directly via leveraging the powerful learning ability of deep neural networks (DNNs). References [15] and [16] propose a DNN-based approach for addressing voltage-constrained AC-OPF problems, while [17] embeds the discrete topology representation into the continuous admittance space to train a DNN for learning the corresponding OPF solution with flexible topology. However, the temporal-coupled devices are not adequately considered in these approaches, which are essentially static OPFs. To incorporate the temporal-coupled devices, [16] presents a convolutional neural network-based approach to coordinate ESSs and further formulate OPF based on the dataset solved as a mixed-integer programming problem. Nonetheless, it is non-trivial and computationally expensive to prepare a comprehensive dataset associated with integer decision variables and is considered NP-hard for large-scale problems. Under this condition, the unsupervised learning paradigm is proposed in [18] to solve OPF. Although the proposed unsupervised learning does not necessitate the ground truths, the weight pertaining to different sub-loss metrics should be carefully designed, and unreasonable assignments could result in violating security constraints or converging into sub-optimal solutions.

RL and its variants have emerged as a promising paradigm for tackling DOPF problems, with deep RL in [19] and safe RL in [20]-[22]. Specifically, [19] introduces penalty-based approaches to enforce action feasibility, but manually setting these penalties can lead to sub-optimal convergence. The Lagrangian-based approaches [20]-[22] are state-of-the-art RL approaches to formulate operational constraints as an augmented Lagrangian function. However, these approaches focus on either continuous action domains or discrete action spaces. Nevertheless, realistic ADN operations involve both discrete actions such as those from on-load tap changers (OLTCs), and continuous actions, including those from PVs and ESSs. The neglect could result in an incompetent ability to simulate ADN operation conditions for deriving OPF. Moreover, the existing RL-based approaches are sensitive to specific topologies within ADNs. Changes in network topology lead to a non-stationary environment, thereby disrupting the stationary assumptions underlying ADN operation. Specifically, the network topology changes cause identical power injections to yield varying power flow outcomes. This variability results in fluctuations in the associated rewards, posing significant challenges in accurately approximating potential returns. Therefore, it is an ongoing topic to integrate the topology information into RL-based approaches for coordinating prevalent discrete-continuous hybrid action domains in ADNs with time-varying network topologies.

In this paper, we endeavor to propose a graph-based safe RL approach for tackling DOPF with hybrid action space in ADNs with time-varying network topologies. The proposed approach consists of three sub-components to simulate DOPF in ADNs with time-varying network topologies: ① parameterized action to model discrete-continuous hybrid action space; ② constrained Markov decision process (CMDP) for modeling the operational constraints in ADN; and ③ embedding graph structure into RL for abstracting topology features. These integrations enable the proposed approach to learn DOPF while accounting for the dynamic, stochastic inherent, and topological characteristics in ADNs. The main contributions are summarized as follows.

1) A novel approach is proposed to implement DOPF in ADNs with time-varying network topologies. In this approach, the graph convolution operator is advocated for parameterized action CMDP on graph to address heterogeneous environments, uncertainties, discrete-continuous action space, and time-coupled devices during the operation of ADNs.

2) The graph convolution operator is integrated into actor-critic networks of the proposed approach. This allows the agent to capture graph-based knowledge from ADN topology and learn the optimal mapping among nodal injections, grid topologies, and DER generations.

3) To deal with safe explorations in ADN with discrete-continuous hybrid action space, primal-dual parameterized action twin delayed deep deterministic policy gradient (PD-PATD3) algorithm is adopted to implement DOPF. The PD-PATD3 employs a hybrid actor-critic network to estimate discrete and continuous action jointly for learning DOPF.

The remainder of this paper is organized as follows. Section II describes the mathematical formulation of DOPF. Section III details the formulation of the parameterized action CMDP on graph for DOPF, while Section IV proposes the graph convolution-based PD-PATD3 for implementing DOPF. Section V presents experimental results on the simplified United Kingdom Generic Distribution System (UKGDS) case. The conclusions are stated in Section VI.

II. Mathematical Formulation of DOPF

We consider an ADN with N nodes denoted by the set of buses $\mathcal{N}=\{1,2,\ldots,N\}$, and the set of branches is represented by $\mathcal{L}$. The set of DERs and the set of loads are denoted as Ω_G and Ω_D, respectively. The objective of DOPF is to find the optimal set-points of DERs and OLTCs to minimize the overall operational cost and adhere to corresponding constraints across multiple time horizons.

A. DERs

The DERs considered in this paper are inverter-based resources. Therefore, they can simultaneously consume or generate active and reactive power through the coordination of inverters. The operation models of ESSs and PVs are introduced in detail below.

1) ESSs

The operational constraints of ESSs are related to the charging/discharging power, energy context, and converter capacity. Thus, the feasible set of ESS is shown in (1).

$$\mathcal{Z}_G^c=\{(p_c,q_c):\underline{p}_d\le p_c\le\bar{p}_c,\underline{E}\le E_t\le\bar{E},p_c^2+q_c^2\le\bar{s}_c^2\} \tag{1}$$

where p_c and q_c are the active and reactive power converted by the ESS converter, respectively; $\underline{p}_d$ and $\bar{p}_c$ are the limits of discharging and changing power, respectively; E_t is the energy context of ESS at time slot t; $\underline{E}$ and $\bar{E}$ are the lower and upper limits of the energy context, respectively; and $\bar{s}_c$ is the capacity of the ESS converter.

The dynamic process of ESSs is shown in (2).

$$E_{t+1}=E_t+(\eta_c[p_{c,t}]^+-\eta_d[p_{c,t}]^+)\Delta t \tag{2}$$

where η_c and η_d are the charging and discharging efficiencies, respectively; $[\cdot]^+$ denotes the $\max(\cdot,0)$; $p_{c,t}$ is the active power converted by the ESS converter at time slot t; and Δt is the interval of the time slot. It is worth noting that the charging and discharging efficiencies are subject to the constraints $\eta_c\le 1$ and $1/\eta_d\le 1$.

2) PV Systems

The operation model of inverter-based PV accounts for the curtailment of active power. Thus, the feasible set of PV is described in (3).

$$\mathcal{Z}_G^{pv}=\{(p_{pv},q_{pv}):0\le p_{pv}\le\bar{p}_{pv},p_{pv}^2+q_{pv}^2\le\bar{s}_{pv}^2\} \tag{3}$$

where p_{pv} and q_{pv} are the active and reactive power from the PV converter, respectively; $\bar{p}_{pv}$ is the maximum active power output of the PV inverter; and $\bar{s}_{pv}$ is the capacity of the PV converter.

B. OLTCs

The OLTC regulates the voltage ratio of an electric transformer by adjusting the turn ratio. The different voltage ratios could result in differentiated power flows and further impact nodal voltage and branch flow across the ADNs. The operation model of OLTC is shown in (4).

$$V=(1+\alpha\cdot Tp_t)V_{sub} \tag{4}$$

where V is the root voltage of ADN; V_{sub} is the rated secondary voltage magnitude of the substation transformer; Tp_t is the position of the OLTC at time slot t; and α is the change ratio per step.

C. Network Model

The AC power flow equations are shown in (5). It is worth noting that AC power flow is a non-convex constraint, rendering a relaxation gap for convex optimization approaches.

$$P_i=\sum_{k=1}^{N}V_iV_k(G_{ik}\cos\theta_{ik}+B_{ik}\sin\theta_{ik}) \tag{5a}$$

$$Q_i=\sum_{k=1}^{N}V_iV_k(G_{ik}\sin\theta_{ik}+B_{ik}\cos\theta_{ik}) \tag{5b}$$

where P_i and Q_i are the net injected active and reactive power at bus i, respectively; G_{ik} and B_{ik} are the real and imaginary elements of the bus admittance matrix, respectively; V_i is the voltage magnitude at bus i; and θ_{ik} is the voltage phase angle difference between bus i and bus k. In scenarios where distribution networks exhibit limited historical data or low observability, a data-driven state estimation algorithm [23] or matrix completion-based state estimation model [24] is employed to estimate the operational state of the distribution network. In such cases, state estimation algorithms are employed to replace the AC power flow equations. This substitution enables the provision of accurate power flow outcomes despite the limited availability of observational data. Besides that, if the prior knowledge about network parameters is unavailable, the surrogate model can be developed for mapping the power injections and power flow, even if with unknown distribution network topology and parameter [25].

D. Objective Function

The objective of DOPF is to minimize the energy purchase cost from the wholesale power market, operational cost of DERs, and OLTC adjustment cost while satisfying the operational constraints of the network and DER. The formulation for DOPF is shown in (6). It is worth noting that the reactive power cost aims to compensate individual generators/bulk systems that provide additional voltage support, which is aligned with economic principles in the competitive market [26].

$$\min\left[\pi_t^pP_t^s+\pi_t^qQ_t^s+\sum_{i\in\Omega_G}(\sigma_i^p|P_{i,t}^G|+\sigma_i^Q|Q_{i,t}^G|)+\sigma^k\Delta k_t\right] \tag{6a}$$

s.t.

$$P_t^s=\sum_{i\in\Omega_D}P_{i,t}^D+\sum_{(i,j)\in\mathcal{L}}I_{ij,t}^2R_{ij}-\sum_{i\in\Omega_G}P_{i,t}^G \tag{6b}$$

$$Q_t^s=\sum_{i\in\Omega_D}Q_{i,t}^D+\sum_{(i,j)\in\mathcal{L}}I_{ij,t}^2X_{ij}-\sum_{i\in\Omega_G}Q_{i,t}^G \tag{6c}$$

$$\underline{V}\le V_{i,t}\le\overline{V} \tag{6d}$$

$$P_{ij,t}^2+Q_{ij,t}^2\le\overline{S}_{ij}^2 \tag{6e}$$

$$(P_{ij,t}-I_{ij,t}^2R_{ij})^2+(Q_{ij,t}-I_{ij,t}^2X_{ij})^2\le\overline{S}_{ij}^2 \tag{6f}$$

$$(1)\text{-}(5) \tag{6g}$$

where P_t^s and Q_t^s are the total active and reactive power imported from the wholesale market at time slot t, respectively; π_t^p and π_t^q are the time-of-use (TOU) prices of active and reactive power at time slot t, respectively; σ_i^p and σ_i^q are the levelized unit operational costs of DERs in terms of active and reactive power, respectively; $P_{i,t}^D$ and $Q_{i,t}^D$ are the aggregated active and reactive power demands of node i at time slot t, respectively; σ^k is the unit per-tap cost of OLTC; $\Delta k_t = \left| Tp_t - Tp_{t-1} \right|$ is the tap position change at time slot t; R_{ij} and X_{ij} are the inverses of G_{ik} and B_{ik}, respectively; $P_{i,t}^G$ and $Q_{i,t}^G$ are the active and reactive power of the DERs at time slot t, respectively; $I_{ij,t}$ is the current at line $(i,j) \in \mathcal{L}$ at time slot t; $P_{ij,t}$ and $Q_{ij,t}$ are the inflow active and reactive power at line $(i,j) \in \mathcal{L}$ at time slot t, respectively, whereas the outflow active and reactive power is denoted as $P_{ij,t} - I_{ij,t}^2 R_{ij}$ and $Q_{ij,t} - I_{ij,t}^2 X_{ij}$, respectively; $V_{i,t}$ is the voltage magnitude at bus i at time slot t; $\underline{V}$ and $\overline{V}$ are the limitations of voltage magnitude; and $\overline{S}_{ij}$ is the thermal constraints of each line. Formula (6a) is the objective function of the DOPF problem. Formulas (6b) and (6c) represent the active and reactive power balance among ADNs. Formula (6d) is the limit of the voltage ranges among the nodes in ADN, while (6e) and (6f) are the complex flow constraints to ensure that the thermal constraints are not violated in both directions of each line in ADN.

III. Formulation of Parameterized Action CMDP on Graph for DOPF

The DOPF is essentially a sequential decision problem, where the distribution network operator (DNO) acts as an agent to interact with ADNs based on the current observations. The goal of RL-based DOPF formulation is to learn a policy that maximizes the cumulative discount reward while minimizing the cumulative discount cost across horizons. To achieve economic power flow management while adhering to operational constraints, the CMDP framework is proposed to simulate the dynamic behavior of DERs and OLTC operations within ADNs. Given that the coordination of DERs and OLTCs involves discrete-continuous hybrid action space, the parameterized action is leveraged to effectively manage this hybrid action space. To further accommodate time-varying network topologies, a graph structure is incorporated into parameterized action CMDP, reformulating it as a parameterized action CMDP on graph, which consists of a tuple $\left\langle \mathcal{S}, \mathcal{H}, \mathcal{T}, \mathcal{R}, \mathcal{C}, \gamma, T \right\rangle$. $\mathcal{S}$ is the state of environments. $\mathcal{H}$ is the hybrid action space. $\mathcal{T}$ is the state transition function to the next state. $\mathcal{R}$ is the reward function given its state and action, while the cost function $\mathcal{C}$ is the penalty. $\gamma \in [0,1]$ is the discount factor. T is a horizon. The main element associated with the parameterized action CMDP on graph is examined as follows.

A. State

To implement the DOPF, the DNO agent makes its decision based on the state of the parameterized action CMDP on graph, including the power injection, operational state of DERs, OLTC position, and power price. These elements can be categorized into two aspects: ① graph-agnostic state, which is represented by a flattened vector in the form $\boldsymbol{s}_t^g = [E_t, \bar{P}_{pv,t}, Tp_t, \pi_t^p, \pi_t^q]$, where $\bar{P}_{pv,t}$ is the maximum PV active power generation; ② graph-based state $\boldsymbol{s}_t^g = [\boldsymbol{P}_t^D, \boldsymbol{Q}_t^D] \in \mathbb{R}^{N \times 2}$, which is denoted as a matrix to represent the active and reactive power demand within ADNs, and $\boldsymbol{P}_t^D = (P_{i,t}^D)$, $\boldsymbol{Q}_t^D = (Q_{i,t}^D)$, $i = 1, 2, \ldots, N$. The graph-agnostic state allows the DNO agent to understand the situation of DERs, OLTCs, and the power market, while the graph-based states enable it to abstract critical features of power demand amid topology changes. These states could orientate appropriate guidance for implementing DOPF under a time-varying network topology. It is worth noting that the levelized operational costs of DERs and OLTCs are assumed to be fixed over the long term, given their service lives of up to 20 years and 40 years, respectively. In scenarios where these costs vary over time, such variations can be integrated into graph-agnostic states to ensure the algorithm remains robust with the generalization capabilities of neural networks.

B. Hybrid Action

The action of the DNO agent contains the discrete action for changing OLTC taps and continuous action for active and reactive power generation of DERs. Thus, it is modeled as a hybrid action space, which is shown in (7).

$$\mathcal{H} = \{(k, \boldsymbol{x}_k) \,|\, \boldsymbol{x}_k \in \mathcal{X}_k, \forall k \in \mathcal{K}\} \tag{7}$$

where $\mathcal{K} = \{0, 1, 2, \ldots, K\}$ is the set of OLTC taps; $\boldsymbol{x}_k = [p_c, q_c, p_{pv}, q_{pv}]$ is the vector of active and reactive power generation of DERs; and $\mathcal{X}_k = \mathcal{Z}_G^c \cup \mathcal{Z}_G^{pv}$ is the feasible operation region of DERs.

C. Reward

Given that the RL-based technique aims to maximize cumulative rewards, while the original DOPF formulation focuses on minimizing operational costs. Under this condition, the snapshot of negative DOPF objective function is adopted as the reward function within the parameterized action CMDP on graph, as shown in (8). It is observed that the reward function can be represented as $\mathcal{R}: \mathcal{S} \times \mathcal{H} \rightarrow \mathbb{R}$, which is a function of the graph-agnostic and graph-based states and the hybrid action. These state and action elements underscore the significance of simultaneously incorporating discrete-continuous hybrid action spaces and ADNs with time-varying network topologies to enable reward formulation for implementing DOPF.

$$r_t = -\pi_t^p P_t^s - \pi_t^q Q_t^s - \sum_{i \in \Omega_G} (\sigma_i^p |P_{i,t}^G| + \sigma_i^Q |Q_{i,t}^G|) - \sigma^k \Delta k_t \tag{8}$$

D. Cost

The cost of the DNO agent includes a set of auxiliary cost functions $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$ to reflect the operational constraints of DERs in (1) and (3), and the operational constraints of ADNs in (6d)-(6f). Under this condition, the cost of the proposed parameterized action CMDP on graph is reformulated as (9), which is denoted as $\mathcal{C}: \mathcal{S} \times \mathcal{H} \rightarrow \mathbb{R}$ to evaluate the immediate cost associated with a state-action pair. It is worth noting that the negative design of cost functions aims to minimize the degree of violating operational constraints.

$$\begin{cases} c_1 := E_t - \overline{E} \le 0 \\ c_2 := \underline{E} - E_t \le 0 \end{cases} \tag{9a}$$

$$c_3 := p_{c,t}^2 + q_{c,t}^2 - \overline{s}_c^2 \le 0 \tag{9b}$$

$$c_4 := p_{pv,t}^2 + q_{pv,t}^2 - \overline{s}_{pv}^2 \le 0 \tag{9c}$$

$$\begin{cases} c_5 := V_{i,t} - \overline{V} \le 0 \\ c_6 := \underline{V} - V_{i,t} \le 0 \end{cases} \tag{9d}$$

$$c_7 := P_{ij,t}^2 + Q_{ij,t}^2 - \overline{S}_{ij}^2 \le 0 \tag{9e}$$

$$c_8 := (P_{ij,t} - I_{ij,t}^2 R_{ij})^2 + (Q_{ij,t} - I_{ij,t}^2 X_{ij})^2 - \overline{S}_{ij}^2 \le 0 \tag{9f}$$

where $p_{c,t}$ and $q_{c,t}$ are the active and reactive power converted by the ESS converter at time slot t, respectively; and $p_{pv,t}$ and $q_{pv,t}$ are the active and reactive power from the PV converter at time slot t, respectively.

E. State Transition

The state transition is denoted as $\mathcal{T}{:}\mathcal{S} \times \mathcal{H} \rightarrow \mathcal{S}$ to specify the transition of the current state to the next state. It is worth noting that both explicit and implicit state transition functions exist in this work. The explicit state transitions have specific physical laws to guide their state, like ESS and OLTC in (2) and (4), respectively. The implicit state transitions include PV generation, TOU price, and power demand, and the proposed approach aims to learn its internal state transition functions as a data-driven paradigm.

The objective of the DNO agent is to learn a policy π: $\mathcal{S} \rightarrow \mathcal{H}$ to maximize the expected discounted cumulative reward while all functions are satisfied, which is shown in (10).

$$\max J(\pi) = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{T} \gamma^t r_t(\boldsymbol{s}_t, k_t, x_{k,t})\right] \tag{10a}$$

$$J_{c_i}(\pi) = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{T} \gamma^t c_{i,t}(\boldsymbol{s}_t, k_t, x_{k,t})\right] \le d_i \quad \forall c_i \in \mathcal{C} \tag{10b}$$

where $J(\pi)$ and $J_{c_i}(\pi)$ are the objective function of expected discounted cumulative reward and cost function, respectively; $\mathbb{E}$ is the expection funcion; $r_t(\boldsymbol{s}_t, k_t, x_{k,t})$ and $c_{i,t}(\boldsymbol{s}_t, k_t, x_{k,t})$ are the reward and cost at time slot t, resepctively; $d_i \ge 0$ is a tolerance parameter, which restricts the violation of constraint (9) within a small value; and $\boldsymbol{s}_t = \{\boldsymbol{s}_t^n, \boldsymbol{s}_t^g\}$, k_t, and $x_{k,t}$ are the state set, discrete action, and continuous action at time slot t, resepctively. This objective reformulates DOPF in an RL framework and is further addressed using a Lagrangian approach. It is worth noting that the RL framework involves multiple periods, which are evaluated using Q-learning and improved through policy gradient to maximize Lagrangian function. The detailed policy evaluation and improvement procedure are presented in Section IV-B.

IV. Graph Convolution-based PD-PATD3 for Implementing DOPF

In this section, a novel graph convolution-based PD-PATD3 is introduced to address DOPF represented by parameterized action CMDP on graph. The proposed approach consists of two components: ① graph convolution-based actor-critic networks; and ② PD-PATD3. Specifically, graph convolution-based actor-critic networks are designed to capture key characteristics among graph-based and graph-agnostic states. The parameters of the aforementioned actor-critic network are updated via the PD-PATD3, ensuring effective policy learning and robust decision-making within dynamic network environments.

A. Graph Convolution-based Actor-critic Networks

The existing actor-critic networks are good at capturing hidden patterns of Euclidean data (e.g., images, text, and videos), but fail to facilitate the non-Euclidean domains represented as graphs. In this work, the graph convolution operator (GCO) is advocated for integration with actor-critic networks to accommodate the power demand under the time-varying network topologies. The proposed framework of graph convolution-based actor-critic networks is shown in Fig. 1.

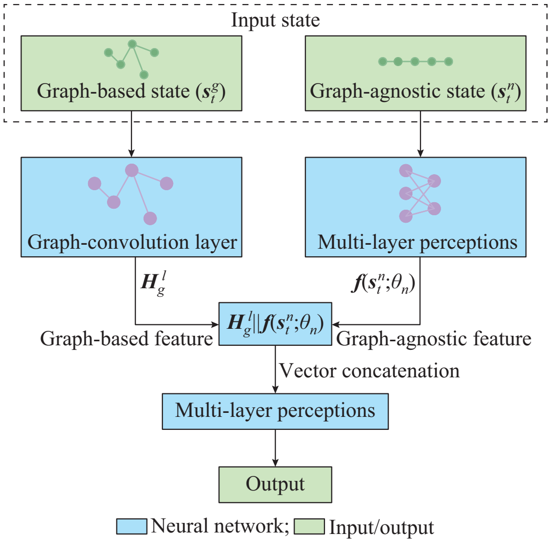
Input state
Graph-based state ($\boldsymbol{s}_t^g$)
Graph-agnostic state ($\boldsymbol{s}_t^n$)
Graph-convolution layer
Multi-layer perceptions
$\boldsymbol{H}_g^l$
$\boldsymbol{f}(\boldsymbol{s}_t^n;\theta_n)$
Graph-based feature
$\boldsymbol{H}_g^l \| \boldsymbol{f}(\boldsymbol{s}_t^n;\theta_n)$
Graph-agnostic feature
Vector concatenation
Multi-layer perceptions
Output
🟦 Neural network; 🟩 Input/output

Fig. 1. Framework of graph convolution-based actor-critic networks.

To abstract the graph-based state of the power demand within ADN, the GCO is advocated to form the normalized Laplacian matrix of the graph on topology. The formulation of GCO is shown in (11).

$$\boldsymbol{G} = \tilde{\boldsymbol{D}}^{-\frac{1}{2}} \tilde{\boldsymbol{A}} \tilde{\boldsymbol{D}}^{-\frac{1}{2}} \tag{11}$$

where $\boldsymbol{G}$ is the graph convolution operator; $\tilde{\boldsymbol{A}}$ is the adjacency matrix of the original graph on ADN topology with self-loops; and $\tilde{\boldsymbol{D}}$ is the diagonal degree matrix of $\tilde{\boldsymbol{A}}$, whose element is denoted as $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. It is worth noting that GCO $\boldsymbol{G} \in \mathbb{R}^{N \times N}$ can be viewed as a graph-based feature based on a specific ADN topology.

Under this condition, the graph state and GCO are fed into the neural network jointly to formulate the graph convolution network (GCN) [27], which is shown in (12).

$$\boldsymbol{H}_g^{l+1} = \sigma(\boldsymbol{G}\boldsymbol{H}_g^l \boldsymbol{W}_g^l) \tag{12}$$

where $\boldsymbol{H}_g^l$ is the entry of layer l within graph convolution layer and it is worth noting that $\boldsymbol{H}_g^0 = \boldsymbol{s}_t^g$ for the input layer; $\boldsymbol{W}_g^l$ is the vector of parameters of graph convolution layer; and σ is the activation function.

The final outputs of the proposed framework of graph convolution-based actor-critic networks are implemented using multilayer perceptions (MLPs). The input data comprise graph-based features abstracted from the GCN and graph-agnostic features abstracted from the MLP, as shown in (13).

$$O = MLP(\boldsymbol{H}_g^l \| \boldsymbol{f}(\boldsymbol{s}_t^n; \theta_n)) \tag{13}$$

where $MLP(\cdot)$ is the MLP function; $\|$ denotes the vector concatenation; $\boldsymbol{f}$ is the MLP embedding function with parameters θ_n; and O is the final output of the proposed framework of graph convolution-based actor-critic networks.

B. PD-PATD3

The proposed PD-PATD3 is a variant of twin delayed deep deterministic policy gradient (TD3) algorithm with the primal-dual approach to constrain a safe exploration and parameterized action to facilitate discrete-continuous hybrid action space [28], [29]. For the primal-dual RL approach, the Lagrangian relaxation procedure is advocated to solve the CMDPs, which is shown in (14).

$$\mathcal{L}(\pi, \lambda) = J(\pi) - \sum_i \lambda_i (J_{c_i}(\pi) - d_i) \tag{14}$$

where $\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ is the set of Lagrangian multipliers. Under this condition, the constrained problem (14) is reformulated as the unconstrained dual problem, which is shown in (15).

$$(\pi^*, \lambda^*) = \arg \min_{\lambda \geq 0} \max_{\pi} \mathcal{L}(\pi, \lambda) \tag{15}$$

where π^* and λ^* are the optimal policy and Lagrangian multiplier, repsectively.

The parameterized action is generated via $\pi(\boldsymbol{s}_t^n, \boldsymbol{s}_t^g)$ and represented by $h_t = \{f(1), f(2), \ldots, f(K), x\}$, where $f(\cdot)$ and x are the representations of discrete action and continuous action, respectively. Under this condition, the target discrete action is denoted as $k = \arg\max_i f(i)$. For evaluating the policy in terms of Q-value, TD3 is introduced to aggregate the state $\boldsymbol{s}_t = \{\boldsymbol{s}_t^n, \boldsymbol{s}_t^g\}$ and hybrid action to calculate the Q-value via the graph convolution-based critic, which is shown in (16).

$$\mathcal{L}_r(\theta_r) = \mathbb{E}_{\mathcal{D}}\left[\left(r_t + \gamma \min_{j=1,2} Q'_{r,j}\left(\boldsymbol{s}_{t+1}, \breve{h}_{t+1}\right) - Q_r(\boldsymbol{s}_t, h_t)\right)^2\right] \tag{16a}$$

$$\begin{cases} \breve{h}_{t+1} = \pi'(\boldsymbol{s}_{t+1}) + \breve{\varepsilon} \\ \breve{\varepsilon} \sim clip(\mathcal{N}(0, \sigma^2), -\varsigma, \varsigma) \end{cases} \tag{16b}$$

where $\mathcal{L}_r(\cdot)$ is the loss function of critic network; $\mathcal{D}$ is the replay buffer of transitions; $\breve{\varepsilon}$ is the policy noise and is clipped by the edge value ς; $\breve{h}_{t+1}$ is the clipped target action; $Q_r(\cdot)$ and $Q'_{r,j}(\cdot)$ are the Q-function and target Q-function with parameters θ_{rj} and θ_{rj}^-, respectively; $\mathcal{N}(0, \sigma^2)$ is the zero-means Gaussan distribution with variance σ^2; $\pi'(\cdot)$ is the target actor; and $clip(\cdot)$ is the clip function with the limitions of $-\varsigma$ to ς.

For estimating the Q-value of the cost function, the cost value function is advocated to estimate its discount cumulative costs. The temporal difference error of the cost function is shown in (17).

$$\mathcal{L}_{c_i}(\theta_{c_i}) = \mathbb{E}_{\mathcal{D}}[(z_{i,t}, -Q_{c_i}(\boldsymbol{s}_t, h_t))^2] \quad c_i \in \mathcal{C} \tag{17}$$

where $z_{i,t} = c_t + \gamma Q'_{c_i}(\boldsymbol{s}_{t+1}, \pi'(\boldsymbol{s}_{t+1}))$ is the target of discount cumulative cost in terms of $c_i \in \mathcal{C}$, and $Q'_{c_i}(\cdot)$ is the target cost function with parameters $\theta_{c_i}^-$; and $Q_{c_i}(\cdot)$ is the cost function with parameters θ_{c_i}.

The actor is updated by applying the policy gradient algorithm to improve the Lagrangian relaxation function $\mathcal{L}(\pi, \lambda)$ regarding the parameters of the actor, which is shown in (18).

$$\nabla_{\theta_\pi} = \mathbb{E}_{\mathcal{D}}[\nabla_{\pi(\theta_\pi)} \mathcal{L}(\pi, \lambda) \nabla_{\theta_\pi} \pi(\theta_\pi)] \tag{18}$$

where ∇ is the gradient operator; and θ_π is the parameter of the actor to generate the policy π based on the state.

The Lagrangian multipliers are updated by using the simple dual gradient ascent, which is shown in (19).

$$\nabla_{\lambda_i} = \mathbb{E}_{\mathcal{D}}[Q_{c_i}(\boldsymbol{s}_t, h_t) - d_i]^+ \quad c_i \in \mathcal{C} \tag{19}$$

The target network is updated via the soft update approach, and the formulation is shown in (20).

$$\theta_{rj} \leftarrow \tau\theta_{rj} + (1-\tau)\theta_{rj}^- \quad j = \{1, 2\} \tag{20a}$$

$$\theta_{c_i} \leftarrow \tau\theta_{c_i} + (1-\tau)\theta_{c_i}^- \quad \forall c_i \in \mathcal{C} \tag{20b}$$

$$\theta_\pi \leftarrow \tau\theta_\pi + (1-\tau)\theta_\pi^- \tag{20c}$$

where τ is the soft parameter for target networks with $\tau \ll 1$; and θ_π^- is the parameter of target actor network.

The pseudo-code of the proposed graph convolution-based PD-PATD3 is shown in Algorithm 1.

Algorithm 1: graph convolution-based PD-PATD3

```
Initialize: $\theta_{r1}$, $\theta_{r2}$, $\theta_{c_i}$, $\theta_\pi$, $\alpha$, $\beta_i$, $d$, $N_l$, and $\mathcal{D}$
for episode changing from 1 to $E$ do
  Reset initial state $\boldsymbol{s}_0$
  for time slot $t$ changing from 1 to $T$ do
    $h_t \sim \pi_{\theta_\pi}(\boldsymbol{s}_t) + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma)$
    Execute hybrid action $h_t$ and get new state $s_{t+1}$, reward $r_t$, and cost $c_t$
    Store $(\boldsymbol{s}_t, h_t, r_t, c_t, \boldsymbol{s}_{t+1})$ in $\mathcal{D}$
    if $N_l$ is larger than the batch size then
      Sample a mini-batch replay buffer $\{\boldsymbol{s}_{t,l}, h_{t,l}, r_{t,l}, c_{t,l}, \boldsymbol{s}_{t+1,l}\}_{l=1}^{N_l}$ from $\mathcal{D}$
      Update $\theta_{r1}$ and $\theta_{r2}$ via minimizing (16)
      Update $\theta_{c_i}$ via minimizing (17)
      if $t$ mod $d$ then
        Perform policy gradient to actor: $\theta_\pi \leftarrow \theta_\pi + \alpha\nabla_{\theta_\pi}$
        Update Lagrangian multipliers via gradient ascent $\lambda_i \leftarrow \lambda_i + \beta_i \nabla_{\lambda_i}$
        Update target network via (20)
      end
    end
  end
end
```

Initially, the DNO agent collects experiences by interacting with the ADNs and adding them to the replay buffer $\mathcal{D}$. If the collected experience instances exceed the minimum batch size N_l, the temporal difference error is formulated to update the parameters of the critic and cost network by minimizing (16) and (17), respectively. The actor is updated via (18) with the policy gradient transmitted from the critic and cost network with learning rate α, while the Lagrangian multipliers are updated via gradient ascent (19) after d steps of delay with learning rate β_i. During the execution stage, the actor directly aggregates graph-agnostic states and graph-based states. It is worth noting that the network models of actor, critic, and cost are all GCN, which can facilitate the proposed graph convolution-based PD-PATD3 to enable DOPF under the time-varying network topologies.

V. Experimental Results

In this section, the experimental results are provided to evaluate DOPF performance in ADNs with time-varying network topologies based on the proposed graph convolution-based PD-PATD3. The algorithms are implemented in Pytorch and run on a PC with Intel$^{(R)}$ Core$^{(TM)}$ i9-10900X CPU.

A. Experimental Setup

The proposed graph convolution-based PD-PATD3 for DOPF is verified via a simplified UKGDS case, where the voltage and branch congestion constraints can be found in [30]. The original network topology is shown in Fig. 2(a) and it is varied every 24 hours. The network topologies for the next 5 days are shown in Fig. 2(b)-(f), respectively. The load profiles are the daily electricity consumptions of 100 low-voltage end users randomly aggregated from [31], while the PV generation profile is sourced from [32]. The technical parameters of ESSs are from [33] and the levelized operational costs of DERs are 3 \$/MWh for active power and 1 \$/Mvarh for reactive power, respectively [34]. The TOU price is selected from [35], while the levelized cost per tap of OLTCs is set to be \$3 [36].

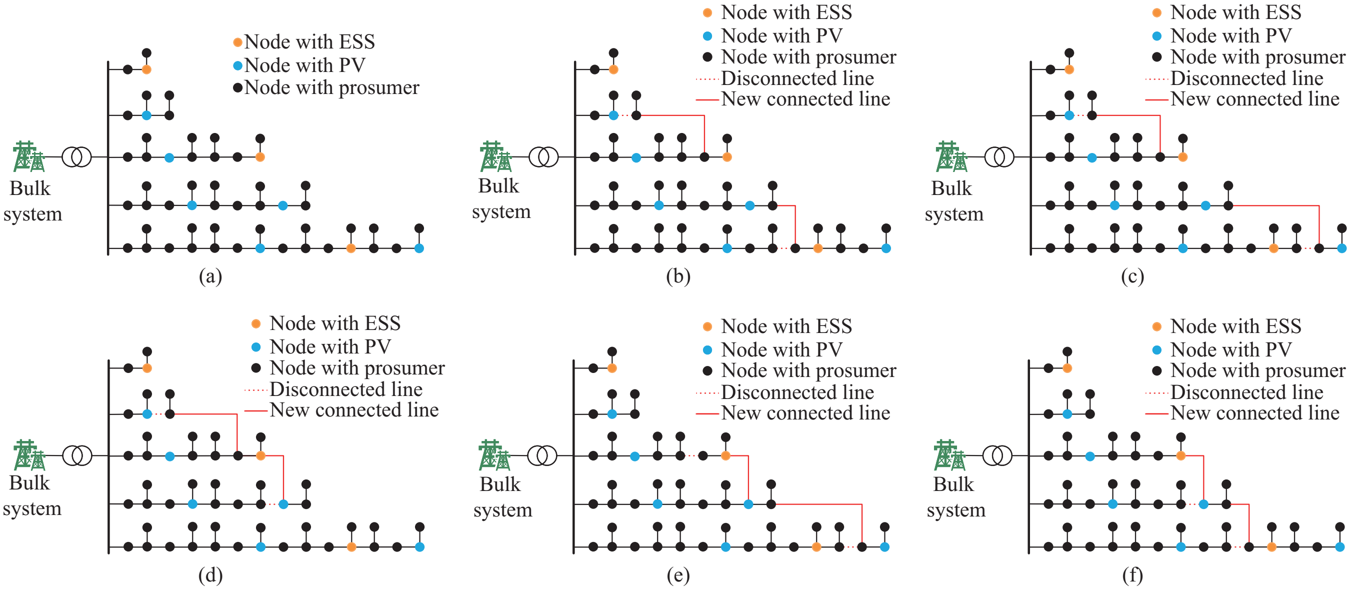

Fig. 2. Original and time-varying network topologies of simplified UKGDS case. (a) Original network topology. (b) Network topology for the 1st day. (c) Network topology for the 2nd day. (d) Network topology for the 3rd day. (e) Network topology for the 4th day. (f) Network topology for the 5th day.

B. Performance Evaluation of Cumulative Rewards

To demonstrate the effectiveness of the proposed approach for implementing DOPF in ADNs, we consider other two graph-based safe RL approaches, that is, the reward-constrained hybrid graph proximal policy optimation (RC-HGPPO) [37] and primal-dual parametrized graph deep Q-network (PD-PGDQN) [38]. The hyperparameters of the proposed approach are summarized as follows. The hidden number of the GCN-based actor-critic networks is 64. The learning rates of Lagrange multipliers, actor, and critic are 5×10^{-5}, 1×10^{-4}, and 2×10^{-4}, respectively. The replay buffer and batch size are set to be 1×10^{6} and 64, respectively. The standard deviations of exploration noise ε and policy noise $\breve{\varepsilon}$ are 0.1 and 0.2, respectively. The discount factor is 0.95, and the soft update parameter is 0.05. The hyperparameters of other two approaches are set to be the same values as those in the proposed approach. The evolution of mean cumulative rewards (line) and associated ranges (shadow) over 5 runs during the training stage is shown in Fig. 3. It is observed that the DNO agent initially receives a low reward because the initialized policy is insufficient to make decisions effectively for DOPF. However, the cumulative reward continuously increases, indicating that the DNO agents have successfully learned a policy to achieve higher rewards. Thanks to the GCN and parameterized action, the proposed approach can tackle non-stationary environments and non-convex mixed-integer programming problems in the ADN with time-varying network topologies. Thus, the proposed approach converges to a higher cumulative reward than the other two approaches and simultaneously achieves the joint exploration of discrete and continuous actions. Furthermore, the RC-HGPPO is an on-policy approach, which has drawbacks in reusing past experiences stored in a replay buffer. This could make it less data-efficient to achieve comparable performance and unstable learning convergence. The overestimation bias of the DQN-based approach results in the PD-PG-DQN encountering challenges in effectively estimating Q-value and stable learning. Furthermore, the oscillations observed during the training phase are attributed to the diverse noise that is introduced to ensure a thorough exploration of the discrete-continuous hybrid action space.

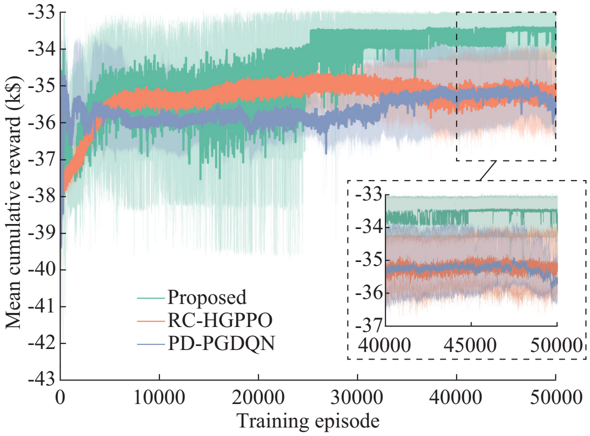

Fig. 3. Evolution of mean cumulative rewards and associated ranges over 5 runs.

The generalization ability of the proposed approach is evaluated on a randomly selected day in terms of cumulative reward and associated standard deviation, as shown in Table I. The simulation results exhibit behavior consistent with the training process, achieving the highest cumulative reward compared with other approaches, thereby highlighting that the proposed approach has the potential to implement DOPF in an unseen environment. Furthermore, it is worth noting that once the proposed approach is trained and the parameters are saved, the actor network can perform the implementation of OPF in real-time, which highlights the efficient application of the RL-based approach in real-life data processing and analysis scenarios.

TABLE I
CUMULATIVE REWARDS AND ASSOCIATED STANDARD DEVIATIONS ON TESTING DAY

Approach	Cumulative reward ($)	Standard deviation ($)
RC-HGPPO	−6981.9	37.26
PD-PGDQN	−7012.5	29.29
Proposed	−6745.5	25.97

C. Performance Evaluation of Cumulative Cost

The cumulative cost is evaluated by accumulating violations of constraints using a Lagrangian multiplier, as shown in (21). This formula is a variant of (10b) to quantify the total penalty for violating operational constraints. The lower cumulative cost indicates that the policy is safer and more compliant with the imposed constraints. The evolution of mean cumulative costs (line) and associated ranges (shadow) over 5 runs during the training stage is shown in Fig. 4.

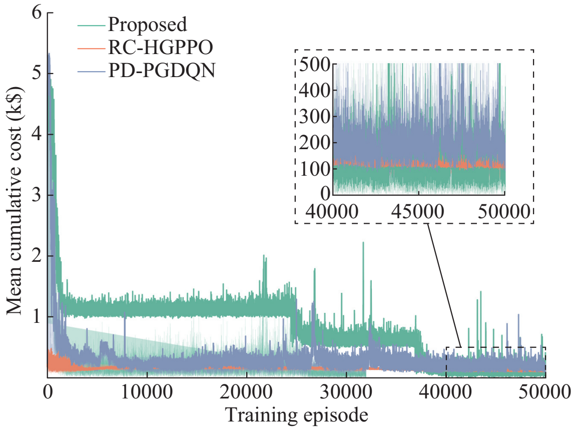

Fig. 4. Evolution of mean cumulative cost and its associated ranges over 5 runs.

$$\mathcal{C}_t = \sum_{c_i \in \mathcal{C}} \lambda_i \, [J_{c_i}(\pi) - d_i]^+ \tag{21}$$

It is observed that the initialization policy yields high cumulative costs, indicating that it is insufficient for safe operation. As the evolution of policy iteration progresses, a more effective policy is derived, resulting in lower cumulative costs. Thanks to the accurate estimation of *Q*-values for state-action pairs provided by the twin critic network and the parameterized action space, the advocated primal-dual mechanism guarantees a safe policy for DOPF within the ADNs.

The generalization ability of the derived safe policy is evaluated on a randomly selected network topology and daily load profile by analyzing the cumulative costs and associated standard deviations, as shown in Table II.

TABLE II
CUMULATIVE COSTS AND STANDARD DEVIATIONS

Approach	Cumulative cost	Standard deviation
RC-HGPPO	14.70	1.75
PD-PGDQN	26.95	2.97
Proposed	4.94	1.61

Specifically, the lower cumulative cost means a safer policy. It is observed that the proposed approach exhibits the lowest cumulative costs compared with other approaches, underscoring the derived safe policy in implementing DOPF under a time-varying network topology.

D. Performance Evaluation of GCN-based Actor-critic Networks

To further evaluate the effectiveness of the GCN-based actor-critic network in handling time-varying network topologies, an ablation study with a fully-connected PD-PATD3 is conducted to emphasize the importance of integrating graph-based structures within the actor-critic framework, which is shown in Table III. The simulation results reveal a 1.73% gap in cumulative reward, but a 609.92% gap is observed in terms of cumulative cost, underscoring the critical role of graph-based structures in neural networks for handling non-stationary environments.

TABLE III
PERFORMANCE EVALUATION ON GCN-BASED ACTOR-CRITIC NETWORKS COMPARED WITH FULLY-CONNECTED PD-PATD3

Approach	Cumulative reward ($)	Cumulative cost
Fully-connected PD-PATD3	−6862.2	35.07
Proposed	−6745.5	4.94

The distribution of the nodal voltage and the apparent power of line under time-varying network topologies is illustrated in Fig. 5. This box plot presents the variation in voltage and apparent power at specific buses or lines over the time horizon. It is observed that the proposed approach effectively manages the nodal voltage and the apparent power of line within safe limits, even under time-varying network topologies. In contrast, the fully-connected PD-PATD3 leads to network feasibility violations, highlighting the importance of integrating graph-based structures in scenarios with topology changes.

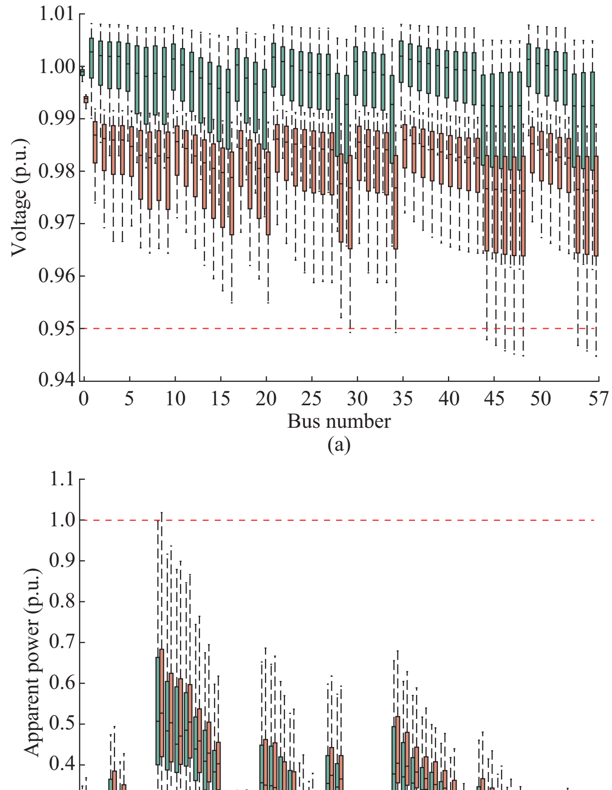

Fig. 5. Distribution of nodal voltage and apparent power of line. (a) Distribution of nodal voltage. (b) Distribution of apparent power of line.

E. Performance Evaluation of Lagrangian-based Safe RL

To evaluate the performance of safe explorations in ADNs with discrete-continuous hybrid action spaces, another ablation study is conducted using the fixed penalty approach. This study aims to highlight the significance of incorporating the Lagrangian relaxation procedure in addressing CMDP, which is shown in Table IV, where ADN constraint adherence means the percentage of not violating operational constraints among all operational constraints. It is worth noting that while the competitors employing fixed penalty values shares a similar mathematical formulation to that of Lagrangian multipliers, the proposed approach offers a straightforward and effective mechanism for determining these Lagrangian multipliers. The simulation results demonstrate that the proposed approach effectively manages the trade-off between cumulative rewards and associated costs while rigorously adhering to operational constraints of ADN. In contrast, the fixed penalty approach either prioritizes maximizing cumulative rewards in scenarios with small penalty values or focuses exclusively on ensuring safe operation in scenarios with large penalty values.

TABLE IV
PERFORMANCE EVALUATION ON LAGRANGIAN-BASED SAFE RL COMPARED WITH FIXED PENALTY APPROACH

Approach	Cumulative reward ($)	Cumulative cost ($)	ADN constraint adherence (%)
Fixed penalty ($\lambda = 0.1$)	−6727.8	37.38	97.1
Fixed penalty ($\lambda = 1$)	−6916.9	22.91	98.2
Fixed penalty ($\lambda = 10$)	−7209.7	7.17	99.7
Fixed penalty ($\lambda = 100$)	−7328.1	2.78	100.0
Proposed	−6745.5	4.94	100.0

VI. CONCLUSION

In this paper, a graph-based safe RL approach is proposed to address the implementation of DOPF in ADNs with time-varying network topologies and hybrid action space. Specifically, a graph convolution-based PD-PATD3 is proposed, which adopts: ① the GCO to tackle non-stationary environments; ② parameterized action for addressing discrete-continuous action spaces; and ③ primal-dual mechanisms to ensure a safe policy. Experimental results demonstrate the effectiveness of the proposed approach for implementing DOPF and superior performance in terms of cumulative reward and cumulative cost compared with other PGSRL approaches. Specifically, the proposed approach shows at least a 3.50% improvement in cumulative reward and a remarkable 197.57% improvement in cumulative cost compared with other graph-based safe RL approaches. The simulation results indicate that the proposed approach enables economically efficient operation while avoiding potential risks of violating operational constraints of ADN.

The potential extension of the proposed work is to: ① explore model-based RL to model spatial-temporal correlations with source-load for improving the sampling efficiency; and ② investigate cyber-physical attack and defense to secure the operation of ADNs.

REFERENCES

[1] M. B. Cain, R. P. Oneill, A. Castillo *et al.*, "History of optimal power flow and formulations," *Federal Energy Regulatory Commission*, vol. 1, pp. 1-36, Dec. 2012.

[2] A. Muhtadi, D. Pandit, N. Nguyen *et al.*, "Distributed energy resourc-

es based microgrid: review of architecture, control, and reliability," *IEEE Transactions on Industry Applications*, vol. 57, no. 3, pp. 2223-2235, May 2021.

[3] F. Nematkhah, F. Aminifar, M. Shahidehpour *et al*., "Evolution in computing paradigms for Internet of Things-enabled smart grid applications: their contributions to power systems," *IEEE Systems*, *Man*, *and Cybernetics Magazine*, vol. 8, no. 3, pp. 8-20, Jul. 2022.

[4] S. Gill, I. Kockar, and G. W. Ault, "Dynamic optimal power flow for active distribution networks," *IEEE Transactions on Power Systems*, vol. 29, no. 1, pp. 121-131, Jan. 2014.

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

[6] Z. Chen, C. Guo, S. Dong *et al*., "Distributed robust dynamic economic dispatch of integrated transmission and distribution systems," *IEEE Transactions on Industry Applications*, vol. 57, no. 5, pp. 4500-4512, Sept. 2021.

[7] E. Craparo, M. Karatas, and D. I. Singham, "A robust optimization approach to hybrid microgrid operation using ensemble weather forecasts," *Applied Energy*, vol. 201, pp. 135-147, Sept. 2017.

[8] M. Lubin, Y. Dvorkin, and L. Roald, "Chance constraints for improving the security of AC optimal power flow," *IEEE Transactions on Power Systems*, vol. 34, no. 3, pp. 1908-1917, May 2019.

[9] M. Vahedipour-Dahraie, H. Rashidizadeh-Kermani, M. Shafie-Khah *et al*., "Risk-averse optimal energy and reserve scheduling for virtual power plants incorporating demand response programs," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1405-1415, Mar. 2021.

[10] Y. Guo, K. Baker, E. Dall'Anese *et al*., "Data-based distributionally robust stochastic optimal power flow: part I: methodologies," *IEEE Transactions on Power Systems*, vol. 34, no. 2, pp. 1483-1492, Mar. 2019.

[11] M. Bazrafshan and N. Gatsis, "Decentralized stochastic optimal power flow in radial networks with distributed generation," *IEEE Transactions on Smart Grid*, vol. 8, no. 2, pp. 787-801, Mar. 2017.

[12] A. Ouammi, H. Dagdougui, L. Dessaint *et al*., "Coordinated model predictive-based power flows control in a cooperative network of smart microgrids," *IEEE Transactions on Smart Grid*, vol. 6, no. 5, pp. 2233-2244, Sept. 2015.

[13] Y. Shi, H. D. Tuan, A. V. Savkin *et al*., "Distributed model predictive control for joint coordination of demand response and optimal power flow with renewables in smart grid," *Applied Energy*, vol. 290, p. 116701, May 2021.

[14] B. Liu, J. H. Braslavsky, and N. Mahdavi, "Linear OPF-based robust dynamic operating envelopes with uncertainties in unbalanced distribution networks," *Journal of Modern Power Systems and Clean Energy*, vol. 12, no. 4, pp. 1320-1326, Jul. 2024.

[15] W. Huang, X. Pan, M. Chen *et al*., "DeepOPF-V: solving AC-OPF problems efficiently," *IEEE Transactions on Power Systems*, vol. 37, no. 1, pp. 800-803, Jan. 2022.

[16] X. Pan, M. Chen, T. Zhao *et al*., "DeepOPF: a feasibility-optimized deep neural network approach for AC optimal power flow problems," *IEEE Systems Journal*, vol. 17, no. 1, pp. 673-683, Mar. 2023.

[17] M. Zhou, M. Chen, and S. H. Low, "DeepOPF-FT: one deep neural network for multiple AC-OPF problems with flexible topology," *IEEE Transactions on Power Systems*, vol. 38, no. 1, pp. 964-967, Jan. 2023.

[18] W. Huang, M. Chen, and S. H. Low, "Unsupervised learning for solving AC optimal power flows: design, analysis, and experiment," *IEEE Transactions on Power Systems*, vol. 39, no. 6, pp. 7102-7114, Nov. 2024.

[19] D. Cao, W. Hu, X. Xu *et al*., "Deep reinforcement learning based approach for optimal power flow of distribution networks embedded with renewable energy and storage devices," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 5, pp. 1101-1110, Sept. 2021

[20] Z. Yan and Y. Xu, "Real-time optimal power flow: a Lagrangian based deep reinforcement learning approach," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 3270-3273, Jul. 2020.

[21] T. Wu, A. Scaglione, and D. Arnold, "Constrained reinforcement learning for predictive control in real-time stochastic dynamic optimal power flow," *IEEE Transactions on Power Systems*, vol. 39, no. 3, pp. 5077-5090, May 2024.

[22] P. Wu, C. Chen, D. Lai *et al*., "Real-time optimal power flow method *via* safe deep reinforcement learning based on primal-dual and prior knowledge guidance," *IEEE Transactions on Power Systems*, vol. 40, no. 1, pp. 597-611, Jan. 2025.

[23] X. Zhang, S. Ge, Y. Zhou *et al*., "Deep learning framework for low-observable distribution system state estimation with multitimescale measurements," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 11, pp. 13273-13283, Nov. 2024.

[24] P. L. Donti, Y. Liu, A. J. Schmitt *et al*., "Matrix completion for low-observability voltage estimation," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2520-2530, May 2020.

[25] J. Yuan and Y. Weng, "Support matrix regression for learning power flow in distribution grid with unobservability," *IEEE Transactions on Power Systems*, vol. 37, no. 2, pp. 1151-1161, Mar. 2022.

[26] D. Jay and K. S. Swarup, "A comprehensive survey on reactive power ancillary service markets," *Renewable and Sustainable Energy Reviews*, vol. 144, p. 110967, Jul. 2021.

[27] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proceedings of International Conference on Learning Representations*, Toulon, France, Feb. 2017, pp. 1-14.

[28] Q. Liang, F. Que, and E. Modiano, "Accelerated primal-dual policy optimization for safe reinforcement learning," in *Proceedings of 31st Conference on Neural Information Processing System*, Long Beach, USA, Feb. 2018, pp. 1-7.

[29] M. Hausknecht and P. Stone, "Deep reinforcement learning in parameterized action space," in *Proceedings of International Conference on Learning Representations*, San Juan, Puerto Rico, Feb. 2016, pp. 1-12.

[30] Centre for Sustainable Electricity and Distributed Generation. (2024, Sept.). United Kingdom generic distribution system. [Online]. Available: https://github.com/sedg/ukgds

[31] A. Koirala, L. Suárez-Ramón, B. Mohamed *et al*., "Non-synthetic European low voltage test system," *International Journal of Electrical Power & Energy Systems*, vol. 118, p. 105712, Jun. 2020.

[32] DKA Solar Centre. (2024, May.). Solar-related knowledge and data from the Northern Territory, Australia. [Online]. Available: https://dkasolarcentre.com.au/

[33] Tesla. (2024, Jul.). Tesla powerwall. [Online]. Available: https://www.tesla.com/en_gb/powerwall

[34] S. A. Newell, T. Andrew, R. Janakiraman *et al*. (2024. Aug.). ERCOT CONE for 2026. [Online]. Available: https://www.brattle.com/wp-content/uploads/2024/08/ERCOT-CONE-for-2026.pdf#page=30.48

[35] PJM. (2024, Aug.). Pennsylvania-New Jersey-Maryland interconnection: markets & operations. [Online]. Available: https://www.pjm.com/markets-and-operations

[36] J. W. Lamont and J. Fu, "Cost analysis of reactive power support," *IEEE Transactions on Power Systems*, vol. 14, no. 3, pp. 890-898, Aug. 1999.

[37] Z. Fan, R. Su, W. Zhang *et al*., "Hybrid actor-critic reinforcement learning in parameterized action space," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, China, Aug. 2019, pp. 2279-2285.

[38] J. Xiong, Q. Wang, Z. Yang *et al*., "Parametrized deep Q-networks learning: reinforcement learning with discrete-continuous hybrid action space," in *Proceedings of International Conference on Learning Representations*, Vancouver, Canada, Feb. 2018, pp. 1-18.

Xihai Zhang is currently pursuing the Ph.D. degree in electrical engineering at Tianjin University, Tianjin, China. His research interests include peer-to-peer energy trading, distribution system state estimation, and reinforcement learning.

Shaoyun Ge received the M.S. degree from the School of Electrical Engineering and Automation, Tianjin University, Tianjin, China, in 1991, and the Ph.D. degree from Hong Kong Polytechnic University, Hong Kong, China, in 1998. He is currently a Professor at the School of Electrical and Information Engineering, Tianjin University. He was a recipient of the State Science and Technology Progress Second Class Award in 2005 and 2010. His research interests include distribution system planning, electric vehicle charging facility planning, and smart grid.

Yue Zhou received the B.Eng. and Ph.D. degrees in electrical engineering from Tianjin University, Tianjin, China, in 2011 and 2016, respectively. He is currently a Professor at the School of Electrical and Information Engineering, Tianjin University. He worked as a Postdoctoral Research Associate and a Lecturer at Cardiff University, Cardiff, UK, during 2017-2020 and 2020-2024, respectively. He led or participated in more than 15 UK and EU projects. He was the CIGRE UK NGN Chair in 2022. His research interests include power system demand side response, electricity market, and cyber-physical system.

Hong Liu received the M.S. and Ph.D. degrees from the School of Electrical Automation Engineering, Tianjin University, Tianjin, China, in 2005 and

2009, respectively. He is currently a Professor at the School of Electrical and Information Engineering, Tianjin University. He was a recipient of the State Science and Technology Progress Second Class Award in 2010 and 2016. His research interests include planning and operation of distribution system and integrated energy system.

Shida Zhang received the B.S. degree from the North China University of Water Resources and Electric Power, Zhengzhou, China, in 2015, the M.S. degree from Hohai University, Nanjing, China, in 2018, and the Ph.D. degree from Tianjin University, Tianjin, China, in 2024, all in electrical engineering. Currently, he is a Research Assistant at the School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou, China. His research interests include distribution system planning and operation.

Changxu Jiang received the Ph.D. degree in electric power system and automation from the School of Electric Power Engineering, South China University of Technology, Guangzhou, China, in 2020. Since 2020, he works at the School of Electrical Engineering and Automation, Fuzhou University, Fuzhou, China. His research interests include coordinated optimization of coupled power-transportation network, smart grid, resilience, stochastic optimization, and deep reinforcement learning.