

Article

Volt/Var Control for Three-Phase Unbalanced Distribution Network Based on Trust Region Safe Reinforcement Learning

Junru Hu ¹, Xiaobo Dou ^{1,*}, Junjie Xiong ² and Xiang Tao ²¹ School of Electrical Engineering, Southeast University, Nanjing 210018, China; 230189639@seu.edu.cn² Electric Power Research Institute, State Grid Jiangxi Electric Power Co., Ltd., Nanchang 330096, China

* Correspondence: dxb_2001@sina.com

Abstract

With the widespread integration of renewable energy, power flow in the system has become extremely complex and variable. This not only exacerbates the operational safety issues of distribution networks but also intensifies the three-phase unbalance situation. The traditional volt/var control (VVC) model is facing significant challenges such as high-dimensional nonlinearity and low efficiency. To address these problems, this paper proposes a volt/var control for three-phase unbalanced distribution network based on trust region safe reinforcement learning. Firstly, a model is constructed based on the three-phase linear power flow matrix. Then it is transformed into a Markov Decision Process (MDP) to overcome the computational burden. Secondly, a trust region construction method based on the Clip mechanism is introduced to ensure stable policy gradient updates and computational efficiency. Further, the Lagrange multiplier is introduced in the trust region optimization to convert the node voltage safety boundary into a cost function, establishing a distribution network safety reinforcement learning (SDRL) model, which limits the output of dangerous action. Finally, through case studies, it is verified that the proposed method can effectively mitigate three-phase unbalance, enhance online decision-making efficiency, and strictly guarantee the safe operation of distribution networks, demonstrating significant feasibility and superiority.

Keywords: volt/var control; three-phase unbalance; active distribution network; safe reinforcement learning; CMDP

1. Introduction

With the strong promotion of environmental protection and clean energy around the world, the scope and quantity of distributed renewable energy sources (RESs) deployment in the power system have been greatly improved. As distribution networks become active distribution networks (ADN), the available control resources within the system have become increasingly abundant. The system's ability to participate in optimization has also been greatly enhanced. However, the output of distributed RESs varies with time, environmental factors which cause situations such as reverse power flow in the network. The possibility of voltage exceeding the limit at the nodes has increased [1]. The distribution network lines have naturally asymmetric parameters, low-voltage users are mainly supplied with single-phase power, and the spatial distribution along the feeder is uneven. The single-phase connection of DG leads to increasingly complex power flow distribution, exacerbating the three-phase imbalance problem [2]. Voltage imbalance in the distribution system may damage components and reduce the efficiency of machines [3].



Academic Editor: José Matas

Received: 1 June 2026

Revised: 22 June 2026

Accepted: 23 June 2026

Published: 29 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Specifically, voltage unbalance induces a current unbalance in induction motors that is 6 to 10 times larger in magnitude, which leads to additional winding losses and overheating. According to the NEMA MG1 standard, when the voltage unbalance reaches 3%, a motor operating close to full load must be derated to approximately 90% of its rated capacity to avoid overheating; once the unbalance exceeds 5%, the temperature rise becomes too rapid for derating alone to provide adequate protection [4]. The operational safety of distribution networks is facing severe challenges.

To mitigate voltage violations caused by three-phase unbalance, Volt/VAR control (VVC) has emerged as a crucial approach. Traditional VVC mainly relies on physical models to coordinate various devices with diverse response times, such as on-load tap changers (OLTCs), static VAR generators (SVGs), and photovoltaic (PV) inverters [5,6]. Conventional approaches predominantly utilize mathematical optimization methods [7,8]. For instance, reference [9] proposed a convex relaxation optimal power flow model based on second-order cone programming to address the nonlinear characteristics of three-phase unbalanced systems, achieving the joint optimization of PV inverters and VVC devices. Reference [10] developed an online voltage control strategy for unbalanced distribution networks based on the projected Newton method. In [11], a VVC scheme based on dynamic cluster partitioning was proposed, which reduces the computational burden through coordinated zonal management. Furthermore, reference [12] achieved multi-source coordinated voltage control via sensitivity approximation based on phasor measurement unit (PMU) data. These traditional mathematical optimization schemes strive to strike a balance between solution accuracy and computational efficiency. However, the power flow equations of three-phase unbalanced distribution networks exhibit high nonlinearity and severe inter-phase coupling. Consequently, the traditional optimization solving process becomes extremely time-consuming and highly reliant on the accuracy of real-time topological parameters. This makes it inadequate to satisfy the millisecond-to-minute-level online real-time regulation requirements under the high penetration of distributed energy resources (DERs). Therefore, there is an urgent need for a model-free control method equipped with ultra-fast online decision-making capabilities.

In recent years, Deep Reinforcement Learning (DRL) has been increasingly applied in the power system domain due to its model-free characteristics and powerful ability to handle high-dimensional nonlinear mappings [13–15]. Reference [16] introduced DRL into VVC, demonstrating its ability to adapt to the dynamic changes of power systems without requiring detailed network models. Various improved DRL algorithms have been successively proposed to further enhance control performance. For instance, reference [17] proposed a Graph-based Proximal Policy Optimization (Graph-PPO) algorithm for voltage optimization, which improved the robustness of the algorithm. The physics-informed surrogate model-based Soft Actor-Critic (PISMBSAC) algorithm proposed in [18] operates independently of network parameters and exhibits robustness against abnormal measurements. Furthermore, reference [19] proposed a DRL-based two-timescale VVC strategy, coordinating multiple devices to accelerate the response speed of online computations. Reference [20] proposed a networked multi-agent Markov Decision Process (MDP) algorithm, which reduces the required data volume and improves communication strategies. However, the aforementioned algorithms face an unavoidable safety bottleneck: due to the lack of strict physical boundary constraints, the random exploration they highly rely on is extremely prone to outputting dangerous actions that trigger voltage violations during optimization or under extreme operating conditions. To address this, some RL approaches introduce penalty functions into the reward function, incorporating the consequences of constraint violations into the learning feedback in the form of operational costs [21–23]. Nevertheless, this “soft constraint” mechanism makes the safety of the distribution net-

work highly dependent on the setting of penalty coefficients. Theoretically, there is still the possibility of limit violations, and it poses a certain risk of causing the non-convergence of policy training.

To address this challenge, researchers have proposed the concept of safe reinforcement learning (Safe RL), which embeds constraints directly into the decision-making process of RL. By accounting for the concept of “safety” before the final action execution, Safe RL not only mitigates the exploration risks associated with control decisions but also ensures convergence to the optimal control policy. Based on whether the embedded constraints are decoupled from the agent’s decision-making process, Safe RL methods can be broadly classified into two categories: decoupled and coupled. Decoupled Safe RL primarily relies on the design of a safety layer. It proactively estimates the safety of the agent’s proposed action and rapidly corrects any actions that violate safety constraints. Reference [24] accomplishes action correction by imitating expert experience, while [25] further introduces a progressively difficult training mechanism to reduce the subjectivity of such experience. Considering the impact of action correction on learning efficiency, references [26,27] project the current action onto the nearest boundary of the safety domain. However, the judgment of action safety heavily relies on the accuracy of the estimations. Furthermore, action correction only targets the safety of the current time step; multiple corrections may be triggered during subsequent agent learning, resulting in prolonged system training times. In contrast to decoupled methods utilizing external safety layers, coupled Safe RL approaches embed safety conditions directly into the policy optimization objective, typically achieved by solving a Constrained Markov Decision Process (CMDP) [28–30]. The advantage of using model-free RL to solve CMDPs lies in its strong universality; it does not rely on precise power flow equations or topological models of the distribution network. Instead, it relies solely on the state trajectories and cost feedback generated through the interaction between the agent and the grid to complete iterative policy optimization [28]. Some papers draw inspiration from TRPO by constructing a trust region based on Kullback–Leibler (KL) divergence, and utilize the cost advantage function to quantify safety constraints. To achieve efficient solving, these methods approximate the objective and constraints using first-order Taylor expansion, and analytically solve the Karush–Kuhn–Tucker (KKT) conditions to adaptively adjust the gradient update direction, theoretically guaranteeing absolute safety during the policy iteration process [29,30].

Nevertheless, the aforementioned optimization process relies on computing the Hessian matrix (or Fisher Information Matrix, FIM) and its inverse, which imposes a tremendous computational burden when dealing with high-dimensional continuous control problems. How to ensure absolute safety during the exploration of high-dimensional three-phase distribution networks while achieving stable and efficient convergence of the control policy—without relying on time-consuming second-order matrix operations—remains a critical technical bottleneck that urgently needs to be resolved.

To address the problems, this paper proposes a volt/var control for three-phase unbalanced distribution network based on trust region safe reinforcement learning. The main contributions are summarized as follows:

1. Build an MDP model for VVC in three-phase unbalanced distribution networks based on the DLF matrix. To deal with the severe inter-phase coupling and asymmetric characteristics of three-phase distribution networks, this paper introduces the three-phase distribution linear flow (DLF) matrix to analytically model the high-dimensional nonlinear power flow features and the physical boundaries of various reactive power regulation devices. Based on this, the complex three-phase VVC problem is reformulated into a standard Markov Decision Process (MDP), effectively overcoming the computational bottleneck caused by traditional power flow calculations in data-driven training.

2. Construct a trust-region policy optimization mechanism based on clipping mechanism to ensure the convergence stability of the control policy. Addressing the issue of training collapse caused by excessively large policy update steps in traditional RL during multi-device coordinated optimization, this paper introduces a trust-region construction method based on clipping mechanism. By limiting the divergence between the old and new policies, this mechanism guarantees monotonic improvement during policy gradient updates, significantly enhancing the optimization efficiency and convergence robustness of the algorithm when dealing with high-dimensional state spaces in three-phase unbalanced systems.
3. Propose a safe deep reinforcement learning (SDRL) algorithm that deeply integrates Lagrange multipliers with the trust-region mechanism. To overcome the problem that traditional “soft constraint (penalty function)” mechanisms fail to prevent voltage violations in distribution networks, this paper introduces the Lagrange multiplier method into the trust-region policy optimization process. By dynamically updating the multipliers, operational limits (e.g., nodal voltages) are transformed into strict constraints. Combined with the clip mechanism, the proposed algorithm not only ensures stable policy updates but also fundamentally restricts the dangerous exploration actions of the agent at the algorithmic level, thereby achieving strict adherence to the safe operational boundaries of the distribution network.

The rest of this paper is organized as follows: Section 2 establishes the VVC model for three-phase unbalanced distribution networks based on the DLF matrix. Section 3 reformulates the three-phase reactive power optimization problem into MDP and describes the trust-region construction method based on the clip mechanism. Section 4 details the trust-region safe reinforcement learning algorithm accounting for the operational constraints of the distribution network, and presents the framework of trust-region SDRL considering operational constraints. Section 5 validates the effectiveness and superiority of the proposed method through simulation case studies. Finally, Section 6 concludes this paper.

2. Model Formulation for Three-Phase Unbalanced Volt/Var Control

2.1. Three-Phase Power Flow Model Based on DLF Matrix

Consider a three-phase radial unbalanced distribution network with $N + 1$ nodes. Relax the connection of the nodes to the main grid, and let $N = 1, 2, \dots, n$ and let $\varepsilon = \{(i, j) | j = p_{ar}(i), i, j \in \mathbb{N}, i \neq j\}$. These two represent the sets of distribution network nodes and lines, respectively. The direct superior node of node j is represented by i .

Figure 1 shows a three-phase line section between bus i and j . The impedance matrix between nodes i and j in Equation (1) follows the standard Carson-formula-based three-phase line model and Kron’s reduction procedure for distribution system analysis [31]; this formulation is directly adopted from the established power engineering literature and is not a contribution of this paper.

$$[Z_{ij}] = \begin{bmatrix} Z_{aa-n}^{ij} & Z_{ab-n}^{ij} & Z_{ac-n}^{ij} \\ Z_{ba-n}^{ij} & Z_{bb-n}^{ij} & Z_{bc-n}^{ij} \\ Z_{ca-n}^{ij} & Z_{cb-n}^{ij} & Z_{cc-n}^{ij} \end{bmatrix} \quad (1)$$

At time t , for any node $i \in N$, the power balance constraint of the three-phase power flow equation is as follows:

$$\begin{cases} P_{i,t}^{\phi, inj} = V_{i,t}^{\phi} \sum_{j=1}^N \sum_{\gamma=a,b,c} V_{j,t}^{\gamma} \left(G_{ij}^{\phi\gamma} \cos \theta_{ij,t}^{\phi\gamma} + B_{ij}^{\phi\gamma} \sin \theta_{ij,t}^{\phi\gamma} \right) \\ Q_{i,t}^{\phi, inj} = V_{i,t}^{\phi} \sum_{j=1}^N \sum_{\gamma=a,b,c} V_{j,t}^{\gamma} \left(G_{ij}^{\phi\gamma} \sin \theta_{ij,t}^{\phi\gamma} - B_{ij}^{\phi\gamma} \cos \theta_{ij,t}^{\phi\gamma} \right) \end{cases} \quad (2)$$

where $\phi, \gamma = a, b, c$ represents the three-phase model; $P_{i,t}^{\phi, inj}$, $Q_{i,t}^{\phi, inj}$ and $V_{i,t}^{\phi}$ represent the net active power and reactive power injected by node i in the ϕ phase and the voltage amplitude of node i ; $V_{j,t}^{\gamma}$ represents the amplitude of the γ phase voltage at node j ; $\theta_{ij,t}^{\phi\gamma}$ represents the phase angle difference between the ϕ phase at node i and the γ phase at node j ; $G_{ij}^{\phi\gamma}$ and $B_{ij}^{\phi\gamma}$ represent the element corresponding to the ϕ phase at node i and the γ phase at node j in the node admittance matrix.

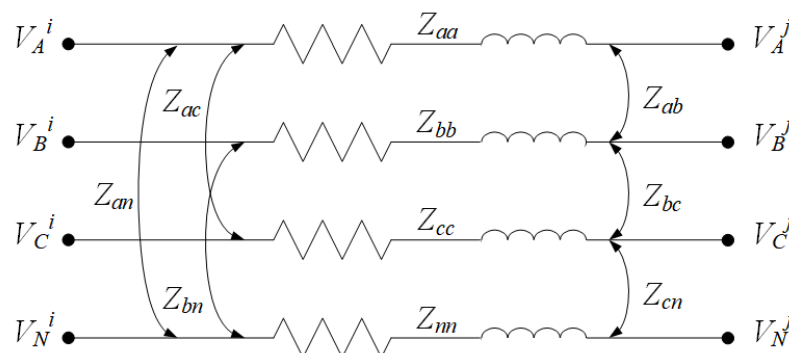


Figure 1. Three-phase line section model.

The traditional Newton-Raphson, forward-backward substitution method requires continuous iterative updates of the Jacobian matrix and LU decomposition when solving the model. This can result in extremely high time costs because of thousands of exploration episodes in reinforcement learning. Teng [32] proposed a power flow solution method based on the DLF matrix for single-phase radial distribution networks, and Equations (1)–(4) extend this single-phase DLF formulation to the three-phase unbalanced case adopted in this paper, in conjunction with the Carson-formula-based three-phase impedance model and Kron's reduction [31]. Based on the topological characteristics of the distribution network, two matrices, BIBC and BCBV, are constructed. Among them, the BIBC matrix represents the relationship between the injected current and the branch current. BIBC is only related to the topological relationship between nodes. The BCBV matrix represents the relationship between the branch current and the bus voltage. It is only related to the line impedance matrix. Thus, these two matrices do not need to be updated in each iteration.

The matrix obtained by multiplying BIBC and BCBV is called the DLF matrix. The relationship between the injected current and voltage can be expressed as:

$$\begin{cases} [\Delta V] = [BCBV][BIBC][I] \\ \quad \quad \quad = [DLF][I] \\ [B] = [BIBC][I] \\ [\Delta V] = [BCBV][B] \end{cases} \quad (3)$$

The three-phase power flow solution based on the DLF matrix is shown in formula 4. The power flow calculation is transformed into pure linear algebraic multiplication. It does not need LU decomposition, and computational complexity is greatly reduced, providing a highly accurate and efficient interactive environment for the following reinforcement learning.

$$\begin{cases} I_i^k = I_i^r(V_i^k) + jI_i^l(V_i^k) = \left(\frac{P_i + jQ_i}{V_i^k} \right)^* \\ \begin{bmatrix} \Delta V^{k+1} \end{bmatrix} = [DLF] I^k \\ \begin{bmatrix} V^{k+1} \end{bmatrix} = \begin{bmatrix} V^0 \end{bmatrix} + \begin{bmatrix} \Delta V^{k+1} \end{bmatrix} \end{cases} \quad (4)$$

where V_i^k and I_i^k represent the bus voltage and equivalent current injection of bus i at the k -th iteration, respectively; I_i^r and I_i^l represent the real and imaginary parts of the equivalent current injection of bus i at the k -th iteration, respectively.

2.2. Controllable Volt/Var Devices' Model

Aiming at volt/var control in three-phase distribution network, this paper focuses on continuous equipment, which mainly includes distributed photovoltaic (PV) and static var generator (SVG). Their output is constrained by the capacity of the equipment and its own output capability:

1. Distributed photovoltaic

PV has both active and reactive power regulation capabilities. This paper considers not reducing the active power output of PV to ensure the complete absorption. The reactive power capacity is adjusted through power factor to provide reactive support for the three-phase system. The model of PV is as follows:

$$\begin{cases} |Q_{PV,t}| \leq P_{PV,t} |\tan \varphi| \\ |\tan \varphi| \leq \tan(\arccos(0.95)) \end{cases} \quad (5)$$

where $\tan \varphi$ represents the minimum power factor of PV.

2. Static var generator

Unlike conventional SVG models that regulate a single three-phase-aggregated reactive power output [9], to enhance the controllability against three-phase asymmetric load disturbances, we model SVGs with phase-independent reactive power injection capability. It allows independent dispatch of Q^a , Q^b , and Q^c for each device, enabling fine-grained mitigation of voltage unbalance caused by single-phase distributed energy resources and asymmetric loads. This per-phase decoupled control structure is illustrated conceptually in Figure 1, where each phase conductor of the SVG branch (terminals A, B, C) can be regulated independently subject to the impedance coupling described in Equation (1), in contrast to a conventional SVG that injects a balanced three-phase reactive power command.

$$Q_i^{SVG,\min} \leq Q_{i,t}^{q,SVG} \leq Q_i^{SVG,\max} \quad (6)$$

where $Q_i^{SVG,\max}$ and $Q_i^{SVG,\min}$ represent the upper and lower bounds of the reactive power compensation for each phase of SVG connected to node i .

It should be clarified that the device-level constraints in Equations (5) and (6) (capacity envelope of PV inverters and the reactive power bounds of SVGs) follow standard models widely adopted in the VVC literature [9,10]. The novel component proposed in this paper is the phase-independent reactive power dispatch of SVGs described below, which is specifically designed to address the three-phase voltage unbalance problem rather than only the conventional positive-sequence voltage regulation considered in most existing single-phase-equivalent VVC formulations.

2.3. Volt/Var Control Model of Three-Phase Distribution Network

Unbalanced distribution of single-phase loads in the distribution network leads to unbalanced three-phase voltages. This may result in increased operating losses, reduced efficiency, and operational limitations. It may even cause damage to electrical components, leading to fault of distribution network. The integration of RES further exacerbates this problem. Experts have proposed the concept of voltage unbalance factor (VUF) to measure effects of negative sequence components on the distribution network:

$$\begin{cases} VUF = \frac{V_2}{V_1} \times 100 \\ V_1 = \frac{1}{3} \left(V^a + \alpha V^b + \alpha^2 V^c \right) \\ V_2 = \frac{1}{3} \left(V^a + \alpha^2 V^b + \alpha V^c \right) \\ \alpha = e^{\frac{2}{3}\pi i} \end{cases} \quad (7)$$

where V_1 and V_2 represent the positive- and negative-sequence components of voltage.

To prevent severe local unbalance issues, a volt/var control model based on min-max objective function is proposed to minimize the worst-case VUF over the entire daily dispatch horizon:

$$\begin{cases} \min F_{obj} = \max_{t \in T} (\max_{i \in \mathcal{N}} VUF_{i,t}) \\ s.t. (5) - (6) \end{cases} \quad (8)$$

where F_{obj} represents the objective function value of the optimization; T is the set of all time steps for the daily scheduling period; $VUF_{i,t}$ indicates the voltage imbalance degree of node i at time t .

3. Trust-Region Based DRL Model for Coordinated Reactive Power Control

Deep reinforcement learning (DRL) adopts a combination of offline training and online operation. During the training episode, it extensively learns the state-action mapping strategy, enabling rapid response in online decision-making. Moreover, DRL also exhibits strong generalization capabilities. Therefore, it has been widely applied in volt/var control.

3.1. MDP Formulation for Unbalanced Three-Phase Distribution System

This paper uses the Markov Decision Process (MDP) to construct the training model for unbalanced voltage mitigation in the distribution network [33]. The dispatch center acts as Agent and the distribution network system serves as Environment. MDP can be described by the tuple (S, A, P, R, γ) :

1. State Space (S)

S represents the current operating status of the power grid. In this paper, it mainly includes: voltage unbalance factor, voltage amplitude at each node, active and reactive power of load, PV and SVG at each node. The state vector is designed to provide the agent with full observability of the quantities that directly determine both the optimization objective and the safety constraints: the VUF and per-phase voltage magnitudes allow the agent to directly perceive the unbalance level and any constraint violation risk at each node, while the active/reactive power of loads, PV, and SVG provide the operating-point information needed to infer the sensitivity of voltage to each control action, given the high inter-phase coupling discussed in Section 2.1.

$$s_{i,t} = (VUF, V_{i,t}^{\varphi}, P_{i,t}^{\varphi,l}, Q_{i,t}^{\varphi,l}, P_{i,t}^{\varphi,PV}, Q_{i,t}^{\varphi,PV}, Q_{i,t}^{\varphi,SVG}) \quad (9)$$

where $s_{i,t}$ represents the state of node i at time t .

2. Action Space (A)

A represents the control measures that agent can take in the current state. In this paper, it mainly includes reactive power of PV inverters and SVGs:

$$\begin{cases} a_{j,t}^{\text{PV}} = \bigcup_{i=1}^n Q_{i,t}^{\varphi,\text{PV}} \\ a_{j,t}^{\text{SVG}} = \bigcup_{i=1}^n Q_{i,t}^{\varphi,\text{SVG}} \end{cases} \quad (10)$$

where $a_{j,t}^{\text{PV}}$ and $a_{j,t}^{\text{SVG}}$, respectively, represent the reactive power output action sets for each phase of PV and SVG. The action space is restricted to continuous reactive power references of PV and SVG, as these devices can respond within seconds and are therefore directly compatible with the millisecond-to-minute control timescale targeted in this paper; discrete and slow-response devices such as OLTCs are outside the scope of the action space and are assumed to operate independently at a slower timescale.

3. Transition Probability (P)

P represents the physical law that, after performing action a_t , the probability of environment transferring to the next state $s_{(t+1)}$. In the power system, the state at $t + 1$ is obtained mainly by solving the equation of the power flow of the power grid. This paper does not need to obtain an explicit state transition probability matrix. Instead, it implicitly learns through the interaction between the power grid simulation environment and the agent. This model-free treatment of P is precisely what distinguishes the proposed framework from the traditional mathematical optimization methods discussed in Section 1: the policy network is updated using only the sampled transitions $(s_t, a_t, r_t, c_t, s_{t+1})$ generated by the DLF-based environment (Section 2.1), without ever requiring P in closed form. This property carries over unchanged when the MDP is extended to the CMDP in Section 4, where the same transition dynamics additionally generate the cost signals C_v and C_{VUF} used to estimate the cost value functions in Equations (20) and (21).

4. Reward Function (R)

Agent learns the optimal policy by maximizing the cumulative reward over the entire dispatch horizon. To align with the optimization objective of minimizing the maximum intra-day global voltage unbalance, the min-max objective is transformed into an immediate reward function at each time step t . Accordingly, the immediate reward r_t is designed as the negative value of the worst (maximum) voltage unbalance factor in the current network, expressed as follows:

$$r_t = -\max_{i \in \mathcal{N}} \text{VUF}_{i,t} \quad (11)$$

5. Discount Factor (γ)

The discount factor is used to measure the conversion ratio of future rewards in the present. This paper takes 0.99 to encourage the intelligent agents to balance the voltage optimization at the current moment and the regulation potential in future periods.

3.2. Trust Region Construction via Clipping Mechanism

Reinforcement learning typically employs the policy gradient method, where the network parameters θ are updated by maximizing objective function $J(\theta) = \mathbb{E}[\log \pi_{\theta}(a|s)\hat{A}]$. The Performance Difference Lemma assumes that the state distribution of the old policy

can be used to evaluate the new policy. However, since three-phase distribution network is high nonlinear and has strong interconnection between phases, when the reactive power control policy undergoes significant changes, the complex power flow can cause huge changes in the system voltage distribution. It will cause a huge high-order error between the approximate evaluation target and the actual grid control performance. If blindly maximizing the advantage function via policy gradient method with an overly large step size, it often causes a local policy collapse. And it will trigger instantaneous voltage violations and deterioration of VUF.

Therefore, in the strategy optimization phase of MDP, this paper introduces a safety mechanism. This paper constructs trust region to restrict the mutation in the mapping from the state space S to the action space A . It constrains the differences between the behavioral (old) and target (new) policies. It strictly limits the update of volt/var control within the safety neighborhood, effectively suppressing the high-order approximation errors caused by the deviation of state distribution, and ensuring that the strategy performance monotonically improvement, which also guarantees the voltage stability of the three-phase power grid during the optimization process.

$$\begin{cases} r_t(\theta) = \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \\ L^{CLIP}(\theta) = \mathbb{E}_t [\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)] \end{cases} \quad (12)$$

where $r_t(\theta)$ represents the ratio of the old and new strategies; $\hat{A}_t$ represents the advantage function; $\text{clip}()$ represents the clip function; ϵ represents the maximum percentage change allowed by the new strategy relative to the old one, usually taken as 0.2.

The traditional algorithm uses the KL Divergence to accurately measure the distance between the old and new strategies to construct the trust region, which causes heavy computational burden. The core concept of the Clip mechanism is to construct an implicit “Algorithmic Trust Region” in the strategy space by limiting the ratio of the old and new strategies. When a certain action brings a positive advantage ($A^t > 0$), the clip function prevents its probability from being unboundedly increased and avoiding overfitting to a single batch of experiences. Conversely, when $A^t < 0$, it limits the quick decrease in the action’s probability, thereby maintaining the exploration potential of the strategy. The first-order clip operation not only prevents huge single-step strategy update, but also avoids the complex FIM matrix calculation. The Clip mechanism perfectly inherits the “anti-collapse” feature of KL divergence while achieving a reduction in computational complexity, as illustrated in Figure 2.

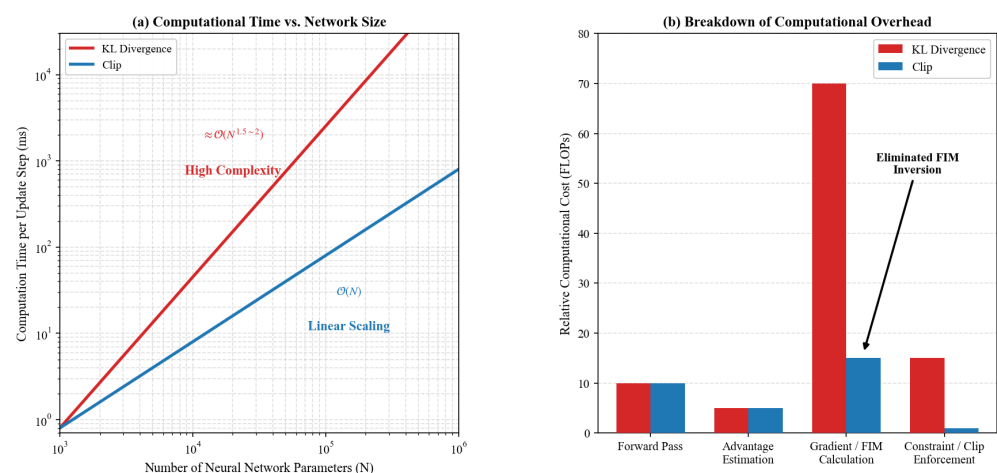


Figure 2. Computational superiority of clip mechanism over KL-divergence.

Meanwhile, $r_t(\theta)$ in the Clip mechanism allows us to conduct multiple epochs of optimization on the same collected rollout before the next round of interaction with the environment. Once the update of the grid control strategy (such as reactive power output, node voltage, etc.) approaches the boundary of the trust region, the Clip mechanism will quickly truncate the gradient to zero. This significantly improves the sample utilization rate of the algorithm and enables the rapid and stable convergence of the grid control strategy.

4. Trust Region Safe Reinforcement Learning Considering Operational Constraints

For reactive power and voltage control in distribution networks, conventional deep reinforcement learning (DRL) faces the dual challenges of insufficient safety guarantees and low convergence efficiency due to its excessive reliance on random exploration. On the one hand, while the action space of DRL is unconstrained, power system operations are subject to many constraints. Although penalty function methods can restrict actions, they are highly sensitive to the tuning of penalty coefficients. In pursuit of economic objectives such as minimizing three-phase unbalance, the exploration process easily triggers action violations, resulting in voltage limit violations. On the other hand, power systems are high-dimensional, complex, and highly sensitive large-scale systems. In the initial stage of training, the agent lacks prior knowledge of environmental dynamics, making it prone to generating extreme actions that can cause severe voltage violations or even system collapse. Therefore, this paper proposes a trust-region safe reinforcement learning method considering operational constraints. By incorporating the Constrained Markov Decision Process (CMDP) and leveraging network topology matrix priors to achieve safe exploration, the proposed method ensures safe system operation while accelerating convergence.

4.1. Constrained Markov Decision Process Considering Power Grid Safety Boundaries

To improve control policy safety, this section transforms distribution network operational constraints into a constrained optimization problem within the RL framework, extending the standard MDP to a Constrained Markov Decision Process (CMDP).

CMDP can be represented by an extended 7-tuple: $(S, A, P, R, C, \gamma, d)$. The safety boundaries are jointly defined by the newly introduced cost function C and the constraint threshold d . Specifically, while maximizing the cumulative reward, the agent must satisfy the expected cumulative cost constraint $\mathbb{E}\left[\sum_{t=0}^T \gamma^t C_t\right] \leq d_c$. In this paper, the cost term C describes the operational constraints of the power system, which specifically fall into the following two categories:

1. Bus voltage safety constraints:

$$C_v = \sum \begin{cases} |v - 1.05| + 0.05, & v > 1.05 \\ |0.95 - v| + 0.05, & v < 0.95 \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

2. Voltage unbalance factor (VUF) safety constraints:

$$C_{VUF} = \sum_{i \in E} \max(0, VUF_i - 0.03) \quad (14)$$

Note that these constraints involve state evolution, which are difficult to restrict directly using simple rules. Thus, cost function is used to guide the learning process considering these constraints. While the reactive power constraints of PVs and SVGs

belong to action-space constraints. They can be directly and strictly satisfied via the activation function mapping in the Actor network.

Consequently, the Volt/Var Control model considering grid safety boundaries can be formulated as follows:

$$\begin{aligned} \max_{\pi} \quad & J_R(\pi) = L^{CLIP}(\theta) \\ \text{s.t.} \quad & J_{C_v}(\pi) = \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t C_v(s_t, a_t) \right] \leq d_v \\ & J_{C_{VUF}}(\pi) = \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t C_{VUF}(s_t, a_t) \right] \leq d_{VUF} \end{aligned} \quad (15)$$

where $J_R(\pi)$ represents the expected cumulative reward of policy π , $J_C(\pi)$ represents the expected cumulative cost corresponding to the cost function, which accounts for the bus voltage safety constraints and the three-phase VUF safety constraints; d_k represents the safety threshold for each constraint.

4.2. Clipped Trust-Region-Constrained Optimization with Lagrange Multipliers

Since the objective of CMDP contains operational constraints, direct solution through policy gradient approaches often suffers from local optimal or training divergence. Therefore, this section introduces the Lagrange multiplier λ to integrate safety into the entire process of policy optimization (PPO-Lag), achieving non-decoupled safe reinforcement learning. By employing the primal-dual method, Equation (17) is transformed into a bi-level optimization problem. A dynamic balance between safety and economic efficiency is then achieved through updating process between the policy network and the dual variable:

$$L(\pi, \lambda) = J_R(\pi) - \lambda(J_{C_k}(\pi) - d_k) \quad (16)$$

$$\max_{\pi} \min_{\lambda \geq 0} L(\pi, \lambda) = L^{CLIP}(\theta) - \sum \lambda_k (J_{C_k}(\pi) - d_k) \quad (17)$$

The Lagrange multiplier is updated as follows:

$$\lambda_k \leftarrow \max(0, \lambda_k + \alpha_{\lambda} (J_{C_k}(\pi) - d_k)) \quad (18)$$

When the action approaches or crosses the trust region boundary, increase λ_k to enhance the penalty weight. If the action is within the trust region, λ_k decreases to release the optimization space of the policy. According to the KKT complementary slackness condition, when reaching the saddle point (π^*, λ^*) , the equation $\lambda_k^* (J_{C_k}(\pi^*) - d_k) = 0$ must be satisfied. Therefore, when achieving Pareto optimality, the policy π^* can strictly satisfy the constraint conditions corresponding to the cost function in the CMDP, and simultaneously maximize the cumulative reward.

Figure 3 illustrates how the clipping mechanism prevents constraint-driven policy oscillation. The plane is divided by the constraint boundary $J_c - d = 0$ into a feasible region ($J_c - d < 0$, left) and an infeasible region ($J_c - d > 0$, right); the optimal point (red star) lies inside the feasible region close to the boundary. The dashed orange trajectory shows the clipped update path: starting from the initial point, each step $k = 1, 2, \dots$ moves the policy parameter θ by $\Delta\theta = \text{Clip}\{\alpha \cdot \nabla_{\theta} [J_R - \lambda(J_C - d)], \pm\epsilon\}$, so that even when the unconstrained gradient direction would push the policy across the boundary (steps $k = 4, 5$ in Figure 3), the clip operation truncates the step size and pulls the trajectory back ($k = 6$: step-back), allowing it to converge smoothly to the optimum by step $k = 8$. In contrast, the dashed purple trajectory shows the same update rule without clipping, $\Delta\theta = \alpha \cdot \nabla_{\theta} J_R$

(with $\lambda = 0$ before any violation is detected): once the trajectory overshoots into the infeasible region (step $k = 5$: overshoot), the dual variable λ increases sharply ($\lambda \gg 1$, step $k = 6$: rebound), producing an oversized corrective gradient that causes the policy to diverge (step $k = 7$: diverge) rather than settle near the optimum. This contrast visualizes why the clipping mechanism introduced in Section 3.2 is necessary for the Lagrangian dual update in Equation (18) to converge stably.

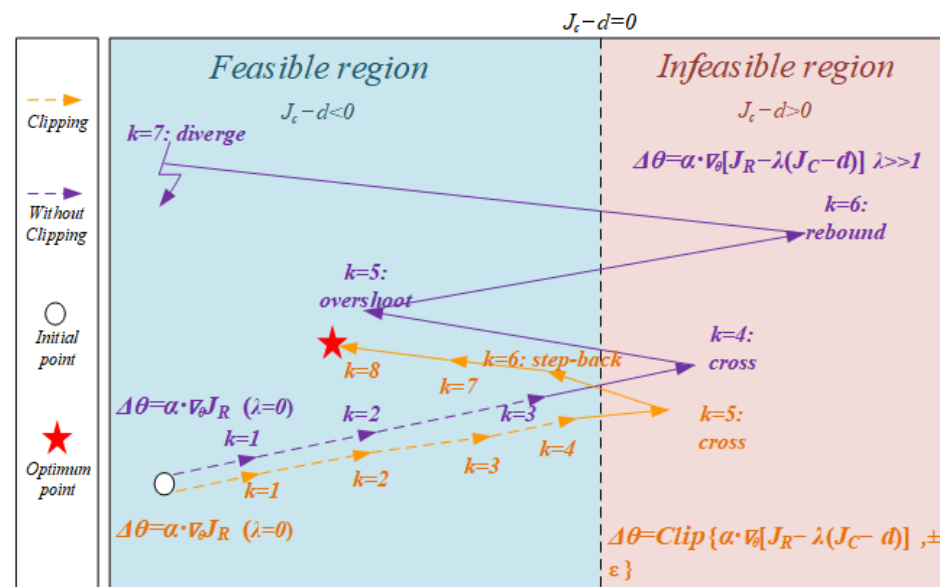


Figure 3. Policy trajectories crossing the constraint boundary, with and without clipping.

Nevertheless, under severe violations of safety constraints (such as severe voltage deviations), the dual variable λ escalates sharply. The resulting volatile updates often trigger severe oscillations in policy gradients, known as gradient spikes. To mitigate this, the clipping mechanism introduced in Section 3.2 constrains the relative magnitude of policy updates within $[1 - \epsilon, 1 + \epsilon]$ through $r_t(\theta)$. As a result, the Lagrange multiplier dictates the update direction, whereas the clipping mechanism ensures a conservative step size. Their cooperation successfully unifies high-sensitivity constraint response with high-stability policy evolution.

4.3. Framework of Trust-Region SDRL Considering Operational Constraints

Framework of trust-region SDRL considering operational constraints (PPO-Lag) is shown in Figure 4. The reinforcement learning framework comprises three core modules: the Environment, the Rollout Buffer, and the Safe Agent.

Environment is responsible for simulating the operating states of the three-phase distribution network. Based on the current state s_t and action a_t , it executes the simulation and returns the immediate reward r_{t+1} , the immediate constraint cost c_{t+1} , and the next state s_{t+1} .

Rollout Buffer is responsible for storing data generated from the interactions between Safe Agent and Environment. Before policy update, it utilizes the collected data to compute the residuals δ_t^R and δ_t^C via the single-step Temporal difference (TD) error estimation method. Subsequently, the advantage functions A_t^R and A_t^C are derived through generalized advantage estimation (GAE). Finally, the reward return target V_t^R and the cost return target V_t^C are reconstructed.

The data flow among the three modules in Figure 4 proceeds as follows. First, the Safe Agent's policy network π_θ samples an action a_t from the current state s_t and sends it to the Environment (top arrow). The Environment executes this action on the three-phase

DLF-based power flow model and returns the tuple $(s_{t+1}, r_{t+1}, c_{t+1})$ to the Rollout Buffer (left arrow), where it is stored together with the value estimates $V_\mu(s_t), V_\mu(s_{t+1})$ and $V_\varphi(s_t), V_\varphi(s_{t+1})$ queried from the current Reward Critic and Cost Critic networks. Once a full episode of T time steps has been collected, the Rollout Buffer computes the TD residuals and GAE-based advantages A_t^R, A_t^C for every stored transition, and passes mini-batches of $(s_t, a_t, A_t^R, A_t^C, V_t^R, V_t^C)$ to the Safe Agent (right arrow) for network updates. Inside the Safe Agent, the same mini-batch is used in parallel to update the three networks: the Actor Loss (Equation (19)) updates the policy network θ using A_t^R, A_t^C , and the current Lagrange multiplier λ ; the Reward Loss (Equation (20)) updates the Reward Critic μ towards V_t^R ; and the Cost Loss (Equation (21)) updates the Cost Critic φ towards V_t^C . The Lagrange multiplier λ is updated separately via Equation (18) using the cost return estimated by the Cost Critic, and is fed back into the Actor Loss for the next iteration, closing the safety feedback loop described in Section 4.2. This cycle (sampling, storage, advantage estimation, and the three coupled network updates) repeats until the policy converges, as detailed in Algorithm 1.

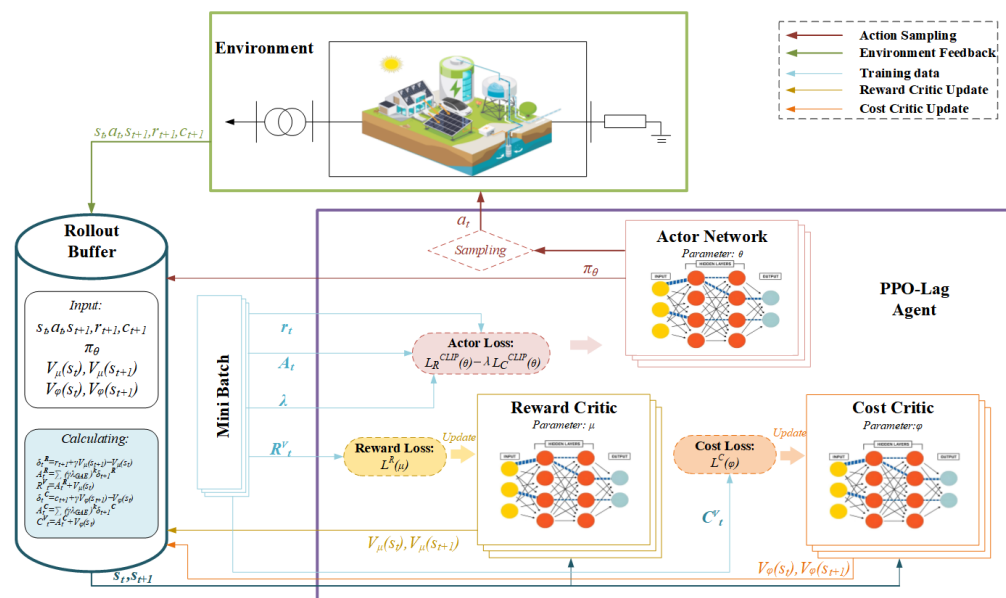


Figure 4. Framework of PPO-Lag.

Safe Agent serves as the core of the algorithm. It retrieves minibatch data from Rollout Buffer to guide the updates of the networks. It employs a multi-network coupled Actor-Critic architecture, achieving policy optimization subject to safety constraints through decoupled evaluation. It contains one policy network and two independent value evaluation networks:

1. Policy network (parameterized by θ). Based on the current state s_t , the policy network outputs a probability distribution over the action space π . Then the action a_t is obtained through a sampling process. Considering safety constraints, the update of the Actor network is transformed into an unconstrained lagrangian dual optimization problem:

$$\max_{\theta} L(\theta, \lambda) = \hat{\mathbb{E}}_t [L^{CLIP}(\theta) - \lambda L_C^{CLIP}(\theta)] \quad (19)$$

2. Reward Critic Network (parameterized by μ). Based on the state s_t , the reward critic network outputs the expected long-term cumulative discounted reward V_μ to evaluate

the long-term benefits of the current control policy in reducing voltage unbalance. The loss function for the network is updated by:

$$L^R(\mu) = \frac{1}{|B|} \sum_{t \in B} \left(V_\mu(s_t) - V_t^R \right)^2 \quad (20)$$

where B represents a randomly sampled mini-batch, and α_μ represents the learning rate for the network update.

3. Cost Critic Network (parameterized by ϕ). Based on the state s_t , the cost critic network outputs the expected long-term cumulative discounted cost V_ϕ to evaluate the potential risk of three-phase voltage limit violations over future time steps. The loss function for the network is updated by:

$$L^C(\phi) = \frac{1}{|B|} \sum_{t \in B} \left(V_\phi(s_t) - V_t^C \right)^2 \quad (21)$$

Algorithm 1 Training procedure of the proposed PPO-Lag algorithm

```

1: Initialize policy network  $\pi_\theta$ , reward critic  $V_\mu$ , cost critic  $V_\phi$ , Lagrange multipliers  $\lambda_k \leftarrow 0$ 
2: for episode = 1, 2, ...,  $N_{\text{ep}}$  do
3:   Reset the three-phase DLF environment to the initial state  $s_0$ 
4:   for  $t = 0, 1, \dots, T - 1$  do
5:     Sample action  $a_t \sim \pi_\theta(\cdot | s_t)$  (reactive power references for PV and SVG)
6:     Apply  $a_t$  to the DLF-based three-phase power flow model, obtain  $s_{t+1}$ , reward  $r_t$  (Equation (11)), and costs  $C_v, C_{VUF}$  (Equations (13) and (14))
7:     Store  $(s_t, a_t, r_t, C_v, C_{VUF}, s_{t+1})$  in the rollout buffer
8:   end for
9:   Compute TD residuals  $\delta_t^R, \delta_t^C$  and advantages  $\hat{A}_t^R, \hat{A}_t^C$  via GAE
10:  for epoch = 1, ...,  $K$  do
11:    Update  $\theta$  by maximizing  $L(\theta, \lambda)$  in Equation (19)
12:    Update  $\mu$  by minimizing  $L^R(\mu)$  in Equation (20)
13:    Update  $\phi$  by minimizing  $L^C(\phi)$  in Equation (21)
14:  end for
15:  Update  $\lambda_k \leftarrow \max(0, \lambda_k + \alpha_\lambda (J_{C_k}(\pi) - d_k))$  for  $k \in \{v, VUF\}$  (Equation (18))
16: end for
17: return Trained policy network  $\pi_\theta$  for online deployment

```

4.4. Training Procedure

The overall offline training procedure of the proposed PPO-Lag algorithm is summarized in Algorithm 1. At each training step, the agent interacts with the three-phase DLF-based environment to collect trajectories, which are used to estimate the reward and cost advantages via GAE, and subsequently to update the policy network, the two critic networks, and the Lagrange multiplier λ_k in turn. This loop is repeated until the policy converges, after which the trained policy network is deployed for online inference without further updates, enabling millisecond-level decision-making.

5. Case Study

5.1. Case Setting

This paper conducts a case study on a standard IEEE123-node three-phase unbalanced system. The system parameters can be found in [30]. PV and SVGs are installed at selected nodes, as shown in Figure 5. All PVs are connected in single-phase and SVGs are connected in three phases with independent per-phase controllability. The detailed configuration parameters are shown in Table 1. Typical residential load and PV generation profiles are

derived from the open dataset provided by the National Renewable Energy Laboratory (NREL) through standardization processing. The profiles are then scaled according to load and PV capacity of the test system, with $\pm 10\%$ random noise capture uncertainty. Then 24-h photovoltaic and load data is achieved.

The SVG installation nodes (48, 74, 76, 82, and 113) are selected based on two principles: proximity to the ends of long feeder laterals, where voltage support is typically weakest and most needed, and proximity to nodes with single-phase PV interconnection, allowing the SVGs to provide localized, phase-independent compensation directly at the source of the three-phase unbalance introduced by the PV units. The listed capacity (500 kvar) represents the single-phase capacity of each independently controllable phase.

It should be clarified that, as shown in Table 1, all PV units are connected in single-phase configuration, while SVGs are connected in three-phase configuration. The capacity values listed for PV in Table 1 (e.g., 600 kW at Node 49, Phase C) therefore represent the single-phase installed capacity.

Table 1. Configuration parameters of PV units and SVG devices.

Type	Capacity	Node	Phase
PV	600 kW	49	C
	600 kW	63	C
	800 kW	75	C
	600 kW	114	A
SVG	500 kvar	48, 74, 76, 82, 113	ABC

Note: A, B, and C denote the three phases of the distribution network; ABC indicates that the device is connected to all three phases independently.

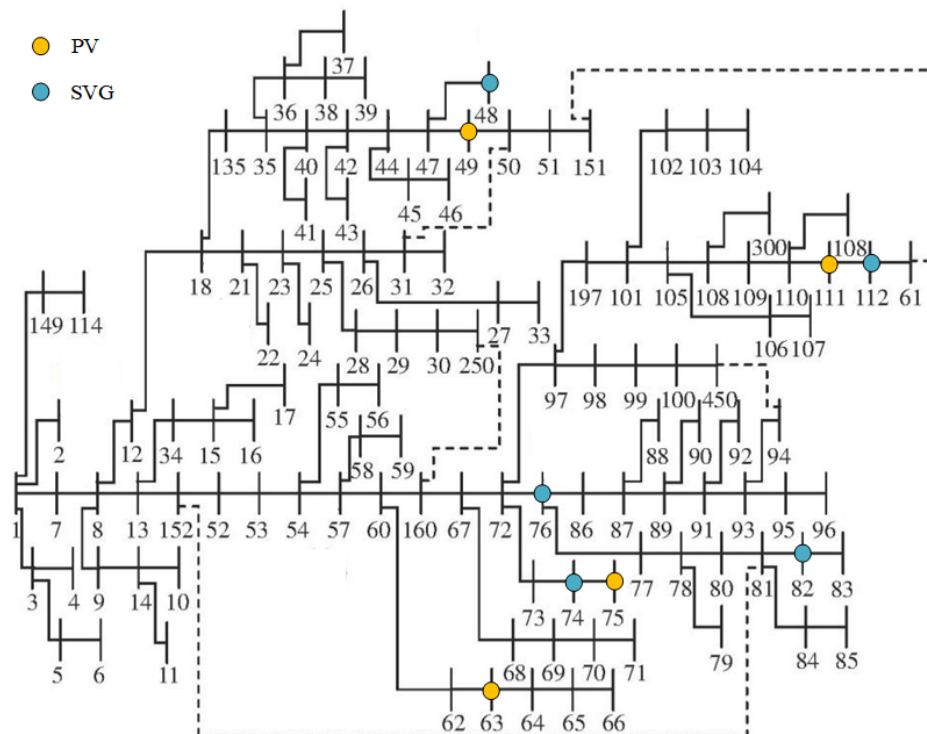


Figure 5. Modified IEEE 123 system.

5.2. Performance in Algorithm Convergence and Collaborative Mechanism

5.2.1. Hyperparameter Settings

The hyperparameter settings used for training the proposed PPO-Lag algorithm, implemented based on the OmniSafe framework, are summarized in Table 2. The clip parameter $\epsilon = 0.2$ follows common practice in the PPO literature and is retained here without retuning, since Section 5.2.2 shows it already yields stable convergence. The discount factor $\gamma = 0.99$ is chosen to be close to 1 so that the agent accounts for the cumulative effect of its reactive power decisions on the VUF trajectory over the full 24 h dispatch horizon, rather than optimizing greedily at each time step. The dual learning rate α_λ is set an order of magnitude larger than the policy/critic learning rates by design, so that the Lagrange multiplier can respond promptly once a constraint violation is detected, while the clipping mechanism discussed in Section 3.2 simultaneously bounds the resulting policy update step within $[1 - \epsilon, 1 + \epsilon]$.

Table 2. Hyperparameter settings of the proposed PPO-Lag algorithm.

Hyperparameter	Value
Policy network learning rate α_θ	3×10^{-4}
Critic network learning rate $\alpha_\mu, \alpha_\varphi$	3×10^{-4}
Dual (Lagrange multiplier) learning rate α_λ	0.035
Clip parameter ϵ	0.2
Target KL divergence	0.05
Discount factor γ	0.99
GAE parameter λ_{GAE}	0.95
Policy/critic network architecture	MLP, 2 hidden layers $\times$ 64 units, tanh activation
Mini-batch size	64
Epochs per update K	10
Training episodes N_{ep}	4000
Voltage safety threshold d_v	0.05 (Equation (13))
VUF safety threshold d_{VUF}	0.03 (Equation (14))

5.2.2. Algorithm Convergence

This paper trains the algorithm for 4000 episodes. Reward convergence curve is presented in Figure 6. It shows that the reward increases steadily in the training process under the conservative update of the clip mechanism. Meanwhile, due to the interaction between the constraint penalty gradient provided by the Lagrange multiplier and the reward gradient, the policy oscillates within a small range near the constraint boundary during the training.

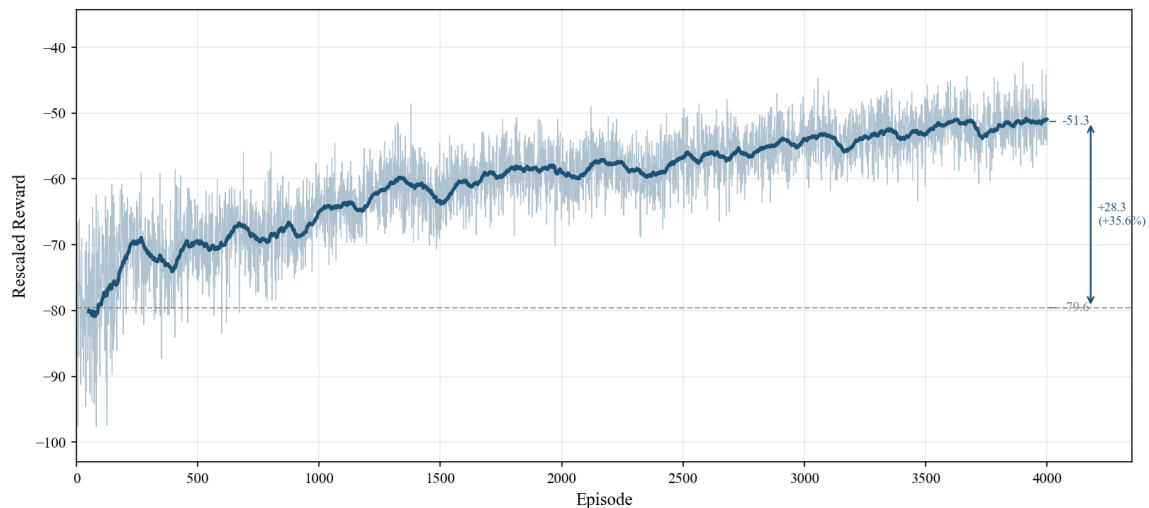


Figure 6. Reward convergence curve of PPO-Lag.

5.2.3. Verification of Collaborative Mechanism

To verify the collaborative mechanism of the clipping mechanism and the Lagrange multiplier in PPO-Lag, TRPO-Lag is additionally trained under identical settings for comparison. The training process of both algorithms are compared in Figure 7.



Figure 7. Comparison of training process between PPO-Lag and TRPO-Lag: (a) Lagrange multiplier λ_k trajectory; (b) Voltage violation cost trajectory.

As shown in Figure 7, the λ_k of TRPO-Lag reaches a peak value of 4.72 and continues to oscillate until around episode 3500, with the mean voltage violation cost remaining at 0.054 in the final training stage. TRPO-Lag's constraint isn't always satisfied during the training process. In contrast, the clipping mechanism in PPO-Lag effectively suppresses the accumulation rate of the dual variable, limiting λ_k peak to 2.39. Then it decreases and converges to zero around episode 1000. Since then, the voltage violation cost of PPO-Lag remains zero, satisfying the KKT complementary slackness condition and confirming that

the policy strictly satisfies the voltage constraints. Results demonstrate that the collaborative mechanism of the clipping mechanism and Lagrange multiplier transforms the oscillatory dual trajectory into a smooth convergence process, achieving constraint satisfaction which TRPO-Lag fails to reach.

5.3. Control Performance Comparison

A typical day is selected for the case study. The voltage heatmap of the system without any reactive power control is presented in Figure 8, where blue indicates $v < 0.95$ and pink indicates $v > 1.05$. To make the safe operating band directly readable from the heatmap itself, the color bar is labeled with its unit (Voltage, p.u.), the horizontal axis is labeled “Hour” with hourly tick marks (H01–H24). The 0.95 p.u. and 1.05 p.u. thresholds referenced in the caption above are marked with dashed reference lines on the color bar, so that violations of either limit can be identified visually without relying solely on the caption text.

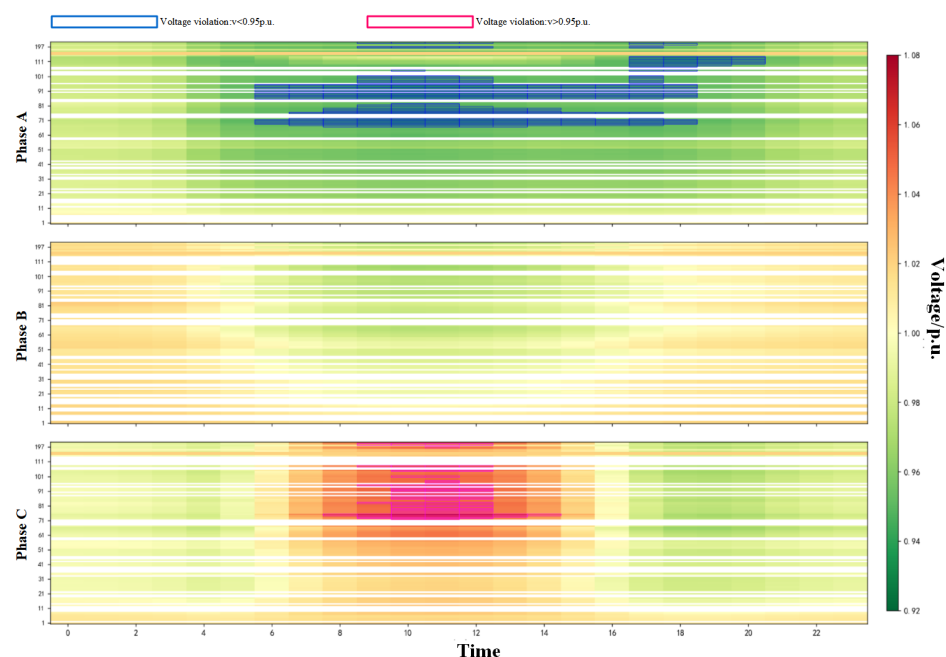


Figure 8. Voltage heatmap of the system without VVC.

It shows that the system experiences severe voltage violation issues during the typical day. During the peak PV generation period (approximately 09:00–15:00), large-scale reverse power flow causes huge voltage rise at nodes near PV. During the peak load periods, insufficient reactive power support leads to voltage drops below 0.95 at nodes located at the end of feeders. Also different phases of the same node exhibit distinctly different voltage profiles at the same time. This is caused by the three-phase imbalance introduced by PV connections, which causes the three phases of the same node to experience fundamentally different voltage conditions. These observations indicate that in three-phase unbalanced distribution networks, voltage violations and three-phase imbalance are mutually coupled. It is necessary to coordinate reactive power resources to simultaneously satisfy voltage constraints and reduce VUF.

5.3.1. VVC Results

Taking 24 time steps as the optimization target, VVC is performed with the objective of minimizing the maximum VUF during the day, subject to voltage constraints. A 24 h VUF comparison is shown in Figure 9.

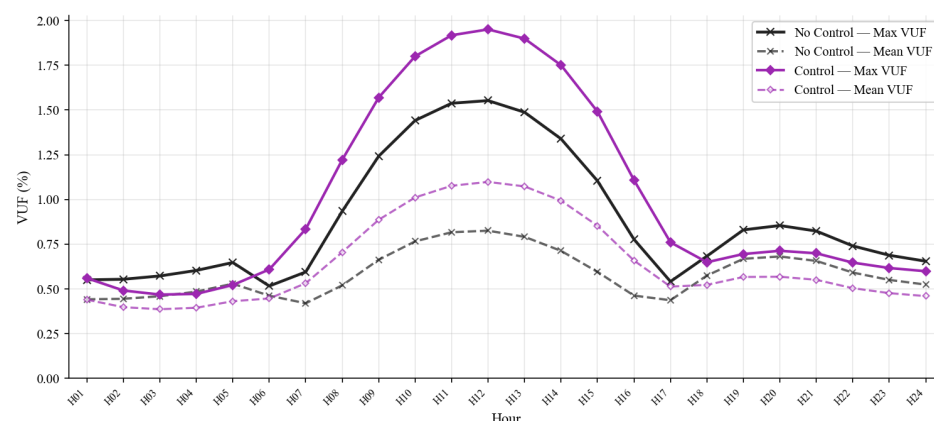


Figure 9. 24-Hour VUF comparison.

As shown in Figure 9, three-phase power imbalance reaches its peak at noon when PV generation is at its maximum. Both the 24 h maximum VUF and mean VUF show a single-peak daily pattern which occurs around noon. When voltage constraint pressure is moderate, VUF after VVC approaches the no-control baseline, demonstrating that the policy effectively optimizes the VUF of the network. During the high PV generation period, both the maximum and mean VUF of the system after VVC exceed the no-control baseline. The gap between the two is slightly larger than that in the no-control case. This is because when system faces voltage violations, λ_k increases, making the policy to prioritize voltage constraint satisfaction at the expense of VUF optimization.

The 24 h reactive power output profiles of the PVs and SVGs are presented in Figure 10.

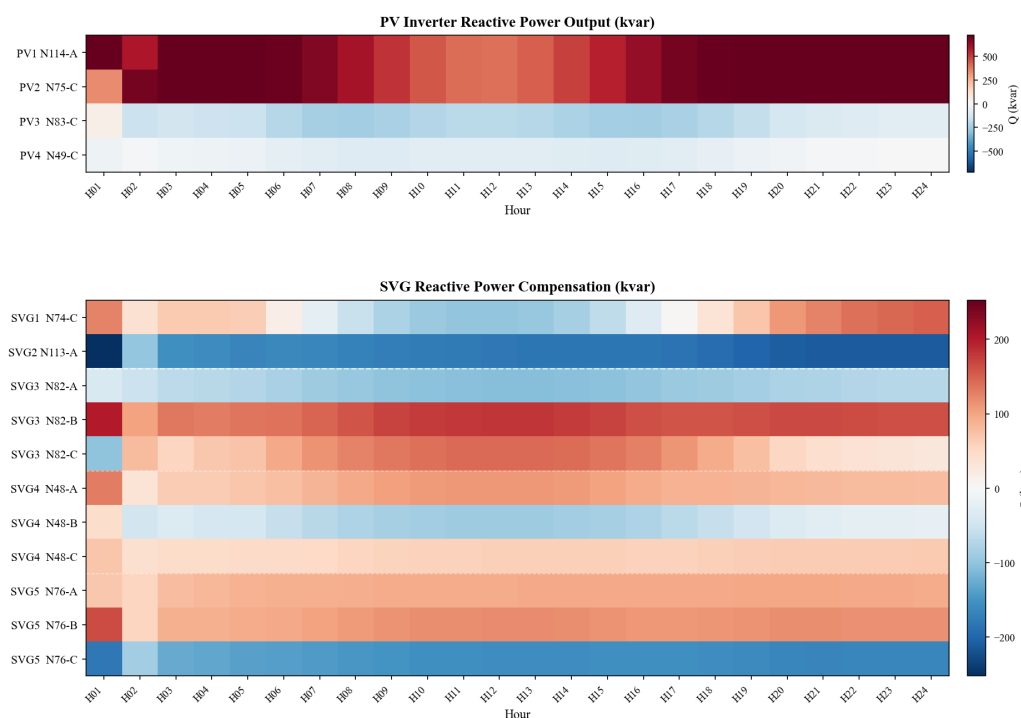


Figure 10. 24 h reactive power output.

As shown in Figure 10, PV inverters perform differentiated reactive power compensation during the peak PV generation period based on the voltage operating conditions. PV1 and PV2 mainly inject capacitive reactive power, while PV3 and PV4 inject inductive reactive power. The SVGs exhibit independent per-phase reactive compensation. Taking SVG3 as an example, the reactive power output of phases A, B and C differs significantly in both

magnitude and direction. Each phase of the SVGs independently performs reactive power compensation, improving three-phase unbalance through coordinated per-phase control.

5.3.2. Comparative Analysis of Algorithms

To verify the superiority of the proposed algorithm in VVC for three-phase unbalanced distribution network, this section compares it against five benchmark schemes: no-VVC baseline, DE global optimization, A2C, PPO, and PPO-Pen(fixed penalty coefficient). Among these, the no-VVC baseline reflects the natural operating state of the system without any control intervention. The comparative analysis is conducted from two perspectives: training characteristics and control performance.

1. Convergence characteristics

The reward convergence curves of the four reinforcement learning algorithms are presented in Figure 11.

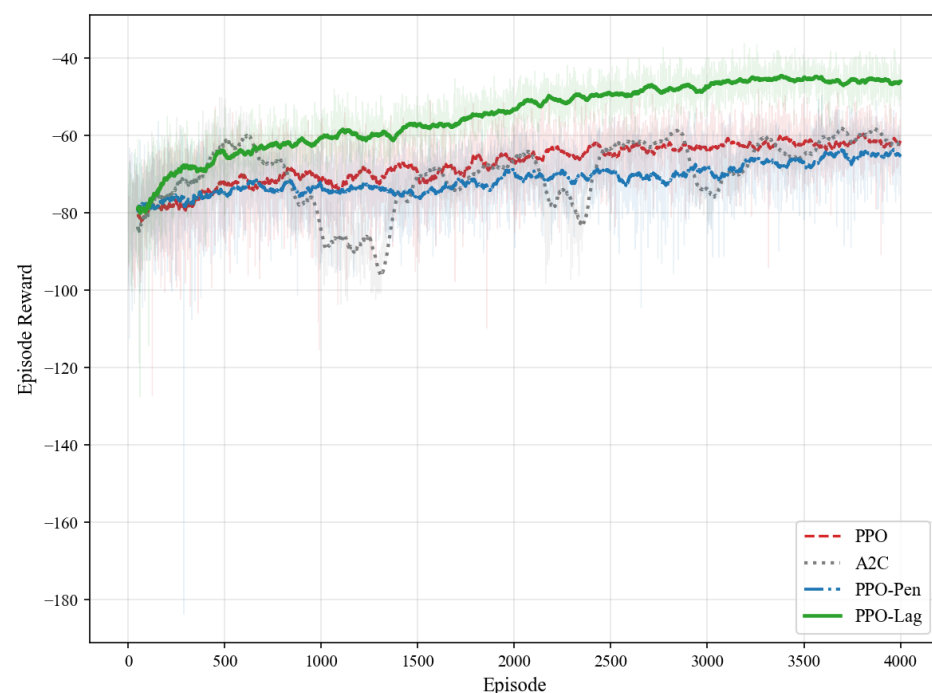


Figure 11. Reward convergence curves of compared algorithms.

A2C employs pure policy gradient updates without any trust region restriction, resulting in the most severe reward oscillations and the highest variance throughout training, making stable convergence difficult to achieve. PPO introduces the clipping mechanism to establish a trust region, achieving noticeably faster convergence and a higher final reward level than A2C under the same unconstrained setting. It shows that the conservative step-size introduced by trust region effectively enhances training stability. PPO-Pen and PPO-Lag inherit the same clipping mechanism, exhibiting reward oscillation amplitudes comparable to PPO and substantially lower than A2C, further confirming that the trust region improves the stability of policy updates.

2. Control performance

Using the same reactive power and voltage control model with minimizing VUF as the objective and node voltage as constraints, the control performance of all algorithms in VVC is summarized in Table 3.

Table 3. Comprehensive performance comparison of all algorithms.

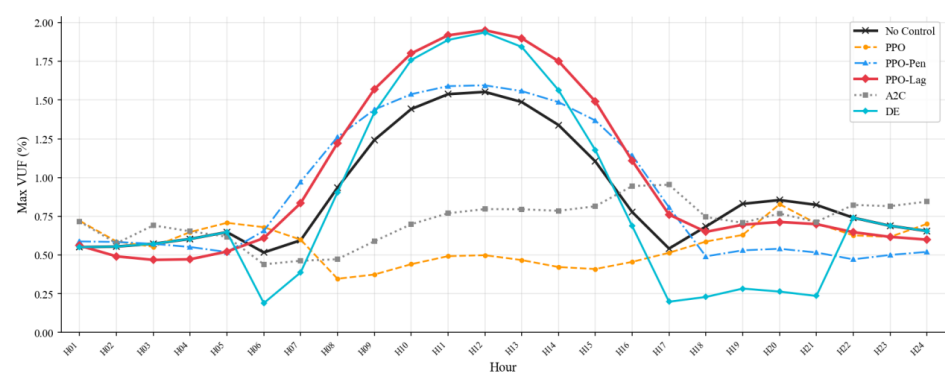
Algorithm	Computation Time (s)		Voltage Violations	Max VUF (%)
	Training	Testing		
No VVC	—		232	1.5523
DE	16,506		0	1.9357
A2C	4483	0.031	384	0.9533
PPO	5155	0.045	432	0.8277
PPO-Pen	5523	0.033	167	1.5947
PPO-Lag	5024	0.021	0	1.9502

Regarding VUF mitigation, the maximum VUF of all algorithms except A2C and PPO—which do not account for voltage constraints—is higher than that of the no-VVC baseline. As shown in Figure 12, the elevated VUF is an inevitable consequence of enforcing voltage safety constraints. It should be noted that the Max VUF of PPO-Lag (1.9502%) remains below the 3% safety threshold d_{VUF} defined in (14), and therefore strictly satisfies the CMDP constraint. According to the NEMA MG1 derating criterion discussed in Section 1, a VUF below 3% only requires a fully loaded motor to be derated to approximately 90% of its rated capacity, which is well within the range that can be safely mitigated through derating rather than posing an imminent overheating risk.

Regarding voltage violations, only PPO-Lag and DE reduce the 24-h count of voltage violation to zero, satisfying the system safety requirements. PPO-Pen compresses the violation count to only 167, but due to insufficient response of the fixed penalty function, the policy cannot be effectively redirected toward the feasible region. Although marginal improvement is achieved within the limited number of training episodes, the safety requirements can not be satisfied.

Regarding computational efficiency, DE requires 16,506 s to finish 24-h optimization. Although PPO-Lag requires 5024 s for training, the actual inference time after training completion is only 0.021 s, representing a substantial improvement in computational efficiency over DE and enabling real-time online deployment.

Thus, PPO-Lag achieves balance between voltage constraint satisfaction and computational real-time performance.

**Figure 12.** Profile of 24 h maximum VUF.

6. Conclusions

This paper proposes a trust-region safe reinforcement learning method for volt/var control in three-phase unbalanced distribution networks. By reformulating the three-phase VVC problem as an MDP based on the DLF matrix, the proposed framework overcomes the computational bottleneck of traditional nonlinear power flow methods. The clipping mechanism constructs an implicit trust region that constrains policy updates within a

conservative step size, while the adaptive Lagrange multiplier dynamically responds to voltage constraint violations through dual ascent. Their cooperative action transforms the oscillatory dual trajectory into a smooth convergence process, reducing the λ_k peak from 4.72 to 2.39 compared with TRPO-Lag and achieving stable constraint satisfaction under identical training parameters.

Case studies on the IEEE 123-node three-phase unbalanced system demonstrate that PPO-Lag is the only scheme that simultaneously achieves zero voltage violations and real-time inference (0.021 s) among all benchmark algorithms. Compared with DE global optimization, PPO-Lag achieves comparable constraint satisfaction with approximately 786,000 times faster online decision-making, validating the engineering practicability and superiority of the proposed method for reactive power and voltage control in three-phase unbalanced distribution networks.

Author Contributions: Conceptualization, J.H. and X.D.; methodology, J.H.; validation, J.H.; resources, J.X. and X.T.; writing—original draft preparation, J.H.; writing—review and editing, J.H. and X.D.; visualization, J.H.; supervision, X.D.; project administration, X.T.; funding acquisition, J.X. and X.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by State Grid Corporation of China, grant number 521800250006-002-ZN.

Data Availability Statement: The simulation data in this study are available from the corresponding author upon reasonable request. The IEEE 123-node test feeder data used in this study are publicly available at <https://site.ieee.org/pes-testfeeders/resources/> (accessed on 10 September 2025).

Acknowledgments: The authors would like to thank the National Renewable Energy Laboratory (NREL) for providing the open-access dataset used in this study.

Conflicts of Interest: Authors Junjie Xiong and Xiang Tao are employed by the Electric Power Research Institute of State Grid Jiangxi Electric Power Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The authors declare that this study received funding from State Grid Jiangxi Electric Power Co., Ltd. (Grant No. 521800250006-002-ZN). The funder was involved in the provision of resources and project administration, but was not involved in the study design, data analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

Abbreviations

The following abbreviations are used in this manuscript:

VVC	Volt/Var Control
DRL	Deep Reinforcement Learning
MDP	Markov Decision Process
CMDP	Constrained Markov Decision Process
DLF	Distribution Load Flow
PPO	Proximal Policy Optimization
SDRL	Safe Deep Reinforcement Learning
VUF	Voltage Unbalance Factor
PV	Photovoltaic
SVG	Static Var Generator
ADN	Active Distribution Network
RES	Renewable Energy Source
DER	Distributed Energy Resource
OLTC	On-Load Tap Changer
KL	Kullback–Leibler

KKT	Karush–Kuhn–Tucker
GAE	Generalized Advantage Estimation
TD	Temporal Difference
FIM	Fisher Information Matrix
TRPO	Trust Region Policy Optimization
PMU	Phasor Measurement Unit
NREL	National Renewable Energy Laboratory
DE	Differential Evolution
A2C	Advantage Actor-Critic

References

1. Zhang, Y.; Wang, Z.; Wang, J. Deep Reinforcement Learning Based Volt-VAR Optimization in Smart Distribution Systems. *IEEE Trans. Smart Grid* **2021**, *12*, 3390–3400. <https://doi.org/10.1109/TSG.2020.3010130>.
2. Ma, K.; Fang, L.; Kong, W. Review of distribution network phase unbalance: Scale, causes, consequences, solutions, and future research directions. *CSEE J. Power Energy Syst.* **2020**, *6*, 479–488. <https://doi.org/10.17775/CSEEJPES.2019.03280>.
3. Pillay, P.; Manyage, M. Definitions of voltage unbalance. *IEEE Power Eng. Rev.* **2001**, *21*, 49–51. <https://doi.org/10.1109/MPER.2001.4311459>.
4. National Electrical Manufacturers Association. *NEMA Standard MG 1: Motors and Generators*; National Electrical Manufacturers Association: Rosslyn, VA, USA, 2021.
5. Wang, C.; Liang, J.; Zhang, L.; Han, X. Volt/VAR Control Strategy for Wind Farms Based on STATCOM. *Proc. CSEE* **2010**, *30*, 23–28. <https://doi.org/10.13334/j.0258-8013.pcsee.2010.25.017>.
6. Huang, Y.; Wang, Y.; Kong, W.; Cao, C.; Su, J.; Wang, K. Active-Reactive Power Coordinated Optimization of Active Distribution Networks Based on Improved Convex Inner Approximation Method. *Proc. CSEE* **2024**, *44*, 5528–5538. <https://doi.org/10.13334/j.0258-8013.pcsee.230267>.
7. Bingane, C.; Anjos, M.E.; Le Digabel, S. Tight-and-Cheap Conic Relaxation for the AC Optimal Power Flow Problem. *IEEE Trans. Power Syst.* **2018**, *33*, 7181–7188. <https://doi.org/10.1109/TPWRS.2018.2848965>.
8. Fu, L.; Liu, B.; Meng, K.; Dong, Z.Y. Optimal Restoration of an Unbalanced Distribution System Into Multiple Microgrids Considering Three-Phase Demand-Side Management. *IEEE Trans. Power Syst.* **2021**, *36*, 1350–1361. <https://doi.org/10.1109/TPWRS.2020.3015384>.
9. He, M.; Soltani, Z.; Ghaljehei, M.; Esmaili, M.; Ma, S.; Chen, M.; Khorsand, M.; Ayyanar, R.; Vittal, V. A SOCP-Based ACOPF for Operational Scheduling of Three-Phase Unbalanced Distribution Systems and Coordination of PV Smart Inverters. *IEEE Trans. Power Syst.* **2024**, *39*, 229–244. <https://doi.org/10.1109/TPWRS.2023.3236881>.
10. Cheng, R.; Wang, Z.; Guo, Y.; Zhang, Q. Online Voltage Control for Unbalanced Distribution Networks Using Projected Newton Method. *IEEE Trans. Power Syst.* **2022**, *37*, 4747–4760. <https://doi.org/10.1109/TPWRS.2022.3144246>.
11. Du, H.; Wei, T.; Xia, D.; Zhou, S.; Han, T. Autonomous-Cooperative Control of Reactive Power and Voltage in Distribution Networks Based on Dynamic Cluster Partitioning. *Autom. Electr. Power Syst.* **2024**, *48*, 171–181.
12. Wang, X.; Wang, C.; Xu, T.; Meng, H.; Li, P.; Yu, L. Distributed voltage control for active distribution networks based on distribution phasor measurement units. *Appl. Energy* **2018**, *229*, 804–813. <https://doi.org/10.1016/j.apenergy.2018.08.042>.
13. Li, Q.; Qiao, Y.; Zhang, Y. Deep Reinforcement Learning Method for Continuous Volt-VAR Optimization in Distribution Networks. *Power Syst. Technol.* **2020**, *44*, 1473–1480. <https://doi.org/10.13335/j.1000-3673.pst.2019.2218>.
14. Zhang, Z.; Zhang, D.; Qiu, R.C. Deep Reinforcement Learning for Power System Applications: An Overview. *CSEE J. Power Energy Syst.* **2020**, *6*, 213–225. <https://doi.org/10.17775/CSEEJPES.2019.00920>.
15. Cao, D.; Hu, W.; Zhao, J.; Zhang, G.; Zhang, B.; Liu, Z.; Chen, Z.; Blaabjerg, F. Reinforcement Learning and Its Applications in Modern Power and Energy Systems: A Review. *J. Mod. Power Syst. Clean Energy* **2020**, *8*, 1029–1042. <https://doi.org/10.35833/MPCE.2020.000552>.
16. Hai, D.; Zhu, T.; Duan, S.; Huang, W.; Li, W. Deep Reinforcement Learning for Volt/VAR Control in Distribution Systems: A Review. In Proceedings of the 2022 5th International Conference on Energy, Electrical and Power Engineering (CEEPE), Chongqing, China, 22–24 April 2022; pp. 596–601. <https://doi.org/10.1109/CEEPE55110.2022.9783357>.
17. Lee, X.Y.; Sarkar, S.; Wang, Y. A graph policy network approach for Volt-Var Control in power distribution systems. *Appl. Energy* **2022**, *323*, 119530. <https://doi.org/10.1016/j.apenergy.2022.119530>.
18. Cao, D.; Zhao, J.; Hu, J.; Pei, Y.; Huang, Q.; Chen, Z.; Hu, W. Physics-Informed Graphical Representation-Enabled Deep Reinforcement Learning for Robust Distribution System Voltage Control. *IEEE Trans. Smart Grid* **2024**, *15*, 233–246. <https://doi.org/10.1109/TSG.2023.3267069>.

19. Li, P.; Jiang, L.; Wang, J.; Xia, H.; Pan, Y. Dual Time-Scale Reactive Power and Voltage Optimization for Renewable Energy Distribution Networks Based on Deep Reinforcement Learning. *Proc. CSEE* **2023**, *43*, 6255–6266. <https://doi.org/10.13334/j.0258-8013.pcsee.221170>.
20. Gao, Y.; Wang, W.; Yu, N. Consensus Multi-Agent Reinforcement Learning for Volt-VAR Control in Power Distribution Networks. *IEEE Trans. Smart Grid* **2021**, *12*, 3594–3604. <https://doi.org/10.1109/TSG.2021.3058996>.
21. Liu, H.; Wu, W. Online Multi-Agent Reinforcement Learning for Decentralized Inverter-Based Volt-VAR Control. *IEEE Trans. Smart Grid* **2021**, *12*, 2980–2990. <https://doi.org/10.1109/TSG.2021.3060027>.
22. Hu, D.; Ye, Z.; Gao, Y.; Ye, Z.; Peng, Y.; Yu, N. Multi-Agent Deep Reinforcement Learning for Voltage Control with Coordinated Active and Reactive Power Optimization. *IEEE Trans. Smart Grid* **2022**, *13*, 4873–4886. <https://doi.org/10.1109/TSG.2022.3185975>.
23. Chen, Z.; Cai, S.; Meliopoulos, A.P.S. Robust Deep Reinforcement Learning for Volt-VAR Optimization in Active Distribution System Under Uncertainty. *IEEE Trans. Smart Grid* **2025**, *16*, 4463–4474. <https://doi.org/10.1109/TSG.2025.3585847>.
24. Peng, Z.; Li, Q.; Liu, C.; Zhou, B. Safe Driving via Expert Guided Policy Optimization. *Conf. Robot Learn.* **2021**, *164*, 1554–1563.
25. Sun, X.; Xu, Z.; Qiu, J.; Liu, H.; Wu, H.; Tao, Y. Optimal Volt/Var Control for Unbalanced Distribution Networks with Human-in-the-Loop Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2024**, *15*, 2639–2651. <https://doi.org/10.1109/TSG.2023.3337843>.
26. Wan, X.; Sun, M.; Chen, B.; Chu, Z.; Teng, F. AdapSafe: Adaptive and Safe-Certified Deep Reinforcement Learning-Based Frequency Control for Carbon-Neutral Power Systems. *AAAI Conf. Artif. Intell.* **2023**, *37*, 5294–5302.
27. Yu, P.; Zhang, H.; Song, Y. District cooling system control for providing regulation services based on safe reinforcement learning with barrier functions. *Appl. Energy* **2023**, *347*, 121396. <https://doi.org/10.1016/j.apenergy.2023.121396>.
28. Zhang, J.; Sang, L.; Xu, Y.; Sun, H. Networked Multiagent-Based Safe Reinforcement Learning for Low-Carbon Demand Management in Distribution Networks. *IEEE Trans. Sustain. Energy* **2024**, *15*, 1528–1545. <https://doi.org/10.1109/TSTE.2024.3355123>.
29. Li, H.; Wang, Z.; Li, L.; He, H. Online Microgrid Energy Management Based on Safe Deep Reinforcement Learning. In Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI), Orlando, FL, USA, 5–7 December 2021; pp. 1–8. <https://doi.org/10.1109/SSCI50451.2021.9659545>.
30. IEEE Power and Energy Society. Distribution Test Feeders. 2010. Available online: <https://site.ieee.org/pes-testfeeders/resources/> (accessed on 10 September 2025).
31. Kersting, W.H. *Distribution System Modeling and Analysis*, 4th ed.; CRC Press: Boca Raton, FL, USA, 2017; 546p. <https://doi.org/10.1201/9781315120782>.
32. Teng, J.H. A direct approach for distribution system load flow solutions. *IEEE Trans. Power Deliv.* **2003**, *18*, 882–887. <https://doi.org/10.1109/TPWRD.2003.813818>.
33. Irfan, M.; Tahir, T.; Deilami, S.; Huang, S.; Veetil, B.P. Novel artificial intelligence driven model-based control framework for solar-electric vehicle home energy optimization. *Eng. Appl. Artif. Intell.* **2026**, *177*, 114868. <https://doi.org/10.1016/j.engappai.2026.114868>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.