Multi-Agent Deep RL for Three-Phase Voltage Control and Balancing in Distribution Systems

Neil Kandukuri
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD, USA

Anqi Liu
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD, USA

Tamim Sookoor
Applied Physics Laboratory
Johns Hopkins University
Laurel, MD, USA

Sijia Geng
Dept. of Electrical & Computer Eng.
Johns Hopkins University
Baltimore, MD, USA

***Abstract*—Voltage control in modern distribution networks is increasingly challenging due to high penetration of distributed energy resources, three-phase unbalance, and higher variability of operating points. We present a zoning-based multi-agent deep reinforcement learning strategy to control voltages using reactive power compensation distributed across the feeder, where voltage unbalance is explicitly treated by employing a Steinmetz-based control as the baseline. The voltage control problem is formulated as a partially observable Markov game, which is then solved by a multi-agent actor-critic deep deterministic policy gradient algorithm with physics-constrained learning objectives. While agents share experiences during training to achieve cooperative voltage regulation, during execution, agents operate in a fully distributed manner without centralized coordination. We demonstrate the effectiveness of the approach on the IEEE 13-bus test feeder with real-world residential load and solar generation data. The RL-informed reactive power compensators were able to eliminate voltage violations while keeping low voltage unbalance across the network.**

***Index Terms*—Voltage control, distribution systems, three-phase unbalance, multi-agent reinforcement learning, policy gradient, distributed control**

I. Introduction

Voltage control is critical for power quality, equipment protection, and reliable service delivery. However, with the increasing adoption of distributed energy resources (DERs), maintaining these limits becomes increasingly difficult for distribution systems as load and generation patterns exhibit high temporal variability and unbalance [1]. Traditional voltage control devices—such as on-load tap changers and switched capacitor banks—operate on slow timescales (minutes to hours) and rely on local measurements, making them infeasible for the fast dynamics introduced by DERs [2]. Conventional phase-balancing methods, such as Steinmetz control, provide circuit-based reactive power compensation to make the equivalent load appears balanced; however, they do not regulate voltage levels [3]. Optimization-based approaches, including optimal power flow and model predictive control, can achieve near-optimal solutions but require accurate system models, centralized computation, and suffer from scalability problems as network size increases [4], [5].

Deep Reinforcement Learning (DRL) methods have demonstrated promising results in power system control applications [6]. DRL agents can learn control policies directly from data without requiring explicit system models, adapt to changing conditions, and handle high-dimensional state-action spaces [7]. Single-agent DRL has been successfully applied to voltage control [8], frequency regulation [9], and energy management [10]. However, single-agent approaches face challenges in large-scale distribution networks due to the need for centralized decision-making and curse of dimensionality.

Multi-agent reinforcement learning (MARL) offers a natural framework for distributed control in networks, where multiple agents can coordinate to achieve system-wide objectives while acting on local information [11]. In power systems, MARL has been explored for microgrid control [12], demand response [13], distribution network reconfiguration [14], and mitigating unbalance [15]. Despite growing interest in learning-based voltage control, several critical gaps remain. First, three-phase unbalance—a fundamental characteristic of distribution systems—is often neglected or not carefully addressed, limiting practical applicability. Second, the impact of network topology on learning-based control performance has not been thoroughly investigated. Third, the trade-offs between voltage regulation performance, computational efficiency, and generalization to unseen scenarios remain poorly understood [16]. Finally, most existing work evaluates performance on synthetic scenarios that may not reflect the diversity and complexity of real-world load and generation patterns. This paper addresses these gaps by proposing a MARL approach for real-time voltage control in unbalanced three-phase distribution systems, by improving upon a baseline provided by the Steinmetz-based balancing control [3]. We design a zone-based multi-agent architecture where agents coordinate implicitly through overlapping observations to regulate voltage across different phases and network zones, enabling scalability. The agents learn physics-informed policies that incorporate power system constraints—including voltage limits, voltage balancing, and reactive power minimization—using real-world residential load profiles and solar generation data.

II. Problem Formulation

A. Radial Distribution Network Model

We consider a general radial three-phase distribution network with $N+1$ buses indexed by $i \in \mathcal{E} = \{1, 2, \ldots, N+1\}$, where bus 1 represents the substation as a slack bus and the rest N buses are load buses. Each bus $i \in \mathcal{E}$ has, without loss of generality, three-phase voltages $V_{i,\phi}$ for phases

Corresponding Author: Sijia Geng, sgeng@jhu.edu.

979-8-3315-8138-1

$\phi \in \{A, B, C\}$, measured in per-unit (p.u.) with appropriate base voltages.

Assume that a subset of J control nodes $\mathcal{J} = \{n_1, n_2, \ldots, n_J\} \subset \mathcal{E}$ are equipped with reactive power compensators for voltage control. The network is partitioned into J control zones, where each zone $\mathcal{Z}_j$ contains buses adjacent to the control node n_j. This zone-based structure enables decentralized control while maintaining system-wide coordination through limited communication requirements.

B. Three-Phase Power Flow and Voltage Unbalance

Denote $\mathbf{v} \in \mathbb{C}^{3N} = [V_{2,A}, V_{2,B}, V_{2,C}, \ldots, V_{N+1,C}]^T \in \mathbb{C}^{3N}$ as the three-phase line-to-neutral voltages at all PQ buses, which can be obtained by solving [17],

$$\mathbf{v} = \mathbf{w} + \mathbf{Y}_{LL}^{-1}\big(\mathrm{diag}(\overline{\mathbf{v}})^{-1}\overline{\mathbf{s}}^{Y} + \mathbf{H}^{\intercal}\mathrm{diag}(\mathbf{H}\overline{\mathbf{v}})^{-1}\overline{\mathbf{s}}^{\Delta}\big), \quad (1)$$

where the overline denotes complex conjugate. $\mathbf{s}^{Y}, \mathbf{s}^{\Delta} \in \mathbb{C}^{3N}$ are the grounded wye- and delta-connected complex power injections at all PQ buses. $\mathbf{w}$ denotes the zero-load voltage profile. $\mathbf{Y}_{LL}$ is the PQ-bus submatrix of the three-phase admittance matrix and $\mathbf{H} \in \mathbb{R}^{3N \times 3N}$ is block diagonal, with $\Gamma = [1, -1, 0; 0, 1, -1; -1, 0, 1]$ on the diagonal and zeros elsewhere, when there are missing phases, the corresponding rows in $\mathbf{H}$ are removed.

The voltage magnitude at each bus-phase must satisfy operational constraints,

$$V_{\min} \le |V_{i,\phi}| \le V_{\max}, \quad \forall i \in \mathcal{E}, \phi \in \{A, B, C\}, \quad (2)$$

where $V_{\min}$ and $V_{\max}$ denote the acceptable voltage range according to utility standards (e.g., ANSI C84.1-2020 specifies $V_{\min} = 0.95$ p.u. and $V_{\max} = 1.05$ p.u.). Violations occur when $|V_{i,\phi}| < V_{\min}$ (undervoltage) or $|V_{i,\phi}| > V_{\max}$ (overvoltage).

The voltage unbalance factor (VUF) quantifies the degree of three-phase unbalance and is defined as the ratio of negative-sequence voltage magnitude V^- to positive-sequence voltage magnitude V^+ [2],

$$\text{VUF} = \frac{|V^-|}{|V^+|} \times 100\%. \quad (3)$$

We aim to control the reactive power injections $\mathbf{Q}_C$ at the J control nodes such that all bus voltages satisfy the constraints in (2) while minimizing VUF, in a distributed manner where each agent operates using only local information.

C. Multi-Agent Markov Game Formulation

We formulate the voltage control problem as a Partially Observable Markov Decision Process (POMDP) game with J cooperative agents, where each agent $j \in \mathcal{J}$ controls reactive power at a control node n_j. Each agent is responsible for regulating voltages within its control zone $\mathcal{Z}_j$ while coordinating with other agents to achieve system-wide objectives. At each time step t, the system transitions from state s_t to s_{t+1} based on the joint action $\mathbf{a}_t = [a_1(t), a_2(t), \ldots, a_J(t)]$ and the stochastic load and generation dynamics.

State Space: Each agent j observes a local observation $o_j(t) \in \mathbb{R}^{d_s}$, where d_s is the dimension of each agent's local observation, consisting of:

- *Local voltages*: Three-phase voltage magnitudes at the agent's control node and nodes in its zone $\mathcal{Z}_j$, zero-padded to ensure uniform dimension across agents.
- *Overlapping voltages*: Voltages from overlapping zones, i.e., buses that are being monitored by two agents, to enable implicit coordination through shared observation.
- *Previous actions*: Reactive power actions from the other agents at the previous step.
- *Global VUF*: A scalar of averaged VUF for all buses in the network.

The global state (used by critics during centralized training) is $s(t) = [o_1(t), o_2(t), \ldots, o_J(t)] \in \mathbb{R}^{J \cdot d_s}$, which concatenates all agents' observations[1]. Note that the global VUF is replicated across observations, providing each agent with system-wide context.

Action Space: Each agent j selects a continuous action $a_j(t) \in \mathcal{A}_j \subseteq \mathbb{R}^{d_a}$ for the reactive power injection at its control node n_j. The actions are bounded to ensure physical realizability: $a_j(t) \in [Q_{\min}, Q_{\max}]^{d_a}$, where $Q_{\min}$ and $Q_{\max}$ define the reactive power capacity limits of the compensators.

Reward Function: The reward function balances local zone performance with global system objectives through a weighted combination. For agent j at time t, the reward is:

$$r_j(t) = w_{\text{local}} \cdot r_j^{\text{local}}(t) + w_{\text{global}} \cdot r^{\text{global}}(t) + r_j^{\text{coord}}(t) + r_j^{\text{act}}(t), \quad (4)$$

where w_{local} and w_{global} are weighting factors that balance local accountability with system-wide performance.

The local reward penalizes voltage violations within the agent's monitored zone $\mathcal{Z}_j$,

$$r_j^{\text{local}}(t) = -\lambda_{\text{local}} \cdot \sum_{k \in \mathcal{Z}_j} \sum_{\phi \in \{A,B,C\}} \mathbb{1}_{[V_{k,\phi} \notin [V_{\min}, V_{\max}]]}, \quad (5)$$

where $\mathbb{1}_{[\cdot]}$ is the indicator function that equals 1 when the voltage is outside acceptable limits, and λ_{local} is a penalty weight.

The global reward incorporates system-wide performance with improvement-based bonuses,

$$r^{\text{global}}(t) = -\lambda_{\text{global}} \cdot N_{\text{viol}}(t) + B_{\text{improve}}(t) + B_{\text{zero}}(t) + B_{\text{VUF}}(t), \quad (6)$$

where λ_{global} is the weighting factor, $N_{\text{viol}}(t)$ is the total number of violations across all buses. $B_{\text{improve}}(t)$ is a tiered bonus for achieving progressive improvement over baseline performance, which provides strong learning signals during training, $B_{\text{zero}}(t)$ and $B_{\text{VUF}}(t)$ reward zero-violation states and VUF reduction.

The coordination bonus encourages compatible actions among agents,

$$r_j^{\text{coord}}(t) = -\alpha \sum_{k \neq j, k \in \mathcal{J}} \|a_j(t) - a_k(t)\|_2, \quad (7)$$

[1]We denote the global state as $s(t) = [o_1(t), \ldots, o_J(t)]$ to emphasize partial observability: each $o_j(t)$ contains only local voltages, previous actions, and the global VUF, not complete grid state (loads, generation, impedances). This POMDP formulation enables distributed execution where agents operate using only local information.

where α is a small weight promoting action similarity for training stability.

The action penalty discourages excessive control effort,

$$r_j^{\text{act}}(t) = -\lambda_{\text{act}} \cdot \|a_j(t)\|_1, \tag{8}$$

where λ_{act} penalizes large reactive power injections.

Objective: The goal is to learn a joint policy $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_J]$ that maximizes the expected cumulative discounted reward,

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\xi\sim\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1}\gamma^t \sum_{j=1}^{J} r_j(t)\right], \tag{9}$$

where $\gamma \in [0,1]$ is the discount factor, T is the episode horizon, and ξ denotes a trajectory sampled from the joint policy. Each individual policy $\pi_j : \mathcal{O}_j \to \mathcal{A}_j$ maps the agent's local observation $o_j(t)$ to its action $a_j(t)$.

III. Multi-Agent Deep RL Methodology

A. Multi-Agent Deep Deterministic Policy Gradient

We employ the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm [18], which extends the DDPG framework [19] to multi-agent settings, to learn coordinated voltage control policies in continuous action spaces. In the actor-critic framework, an *actor* network learns a policy mapping observations to actions, while a *critic* network evaluates action quality. MADDPG extends the actor-critic framework to multi-agent settings through a centralized training with distributed execution (CTDE) paradigm. Each agent maintains four neural networks:

- *Actor network* $\mu_{\theta_j} : \mathcal{O}_j \to \mathcal{A}_j$ that maps local observations to continuous actions.
- *Critic network* $Q_{\phi_j} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ that evaluates value functions, that is, the total expected reward from the current time step until the end of the time horizon, using global state and joint actions of all agents.
- *Target actor network* $\mu_{\overline{\theta}_j} : \mathcal{O}_j \to \mathcal{A}_j$ for stable training.
- *Target critic network* $Q_{\overline{\phi}_j} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ for stable training.

Here, θ_j and ϕ_j denote the parameters of agent j's actor and critic networks, while $\overline{\theta}_j$ and $\overline{\phi}_j$ are parameters for the iteratively updated target networks. These parameters are initialized randomly, and updated during training based on the following steps.

The critic Q_{ϕ_j} is trained by minimizing the temporal difference (TD) error,

$$\mathcal{L}_{critic}(\phi_j) = \mathbb{E}_{(s,\mathbf{a},r,s')\sim\mathcal{D}}\left[\left(Q_{\phi_j}(s,\mathbf{a}) - y_j\right)^2\right], \tag{10}$$

where s' is the next global state after executing joint action $\mathbf{a}$ in state s, and $\mathcal{D}$ is the shared multi-agent experience replay buffer storing transitions. Each transition stores the joint observation, joint actions, individual rewards, next joint observation, and termination flags for all agents. During training, we sample mini-batches from $\mathcal{D}$ and perform gradient updates for all agents simultaneously. The TD target y_j is computed using target networks,

$$y_j = r_j + \gamma Q_{\overline{\phi}_j}(s', \mu_{\overline{\theta}_1}(o_1'), \ldots, \mu_{\overline{\theta}_J}(o_J')), \tag{11}$$

where $\gamma \in [0,1]$ is the discount factor that weights future rewards, $o_j' \in \mathbb{R}^{d_s}$ is agent j's local observation at the next time step.

The actor is updated by maximizing the expected return using the deterministic policy gradient,

$$\nabla_{\theta_j} J(\theta_j) = \mathbb{E}_{s\sim\mathcal{D}}\big[\nabla_{\theta_j}\mu_{\theta_j}(o_j)\cdot \nabla_{a_j} Q_{\phi_j}(s, a_1, \ldots, a_J)\big|_{a_j=\mu_{\theta_j}(o_j)}\big]. \tag{12}$$

Target networks are updated via soft updates,

$$\overline{\theta}_j \leftarrow \tau\theta_j + (1-\tau)\overline{\theta}_j, \quad \overline{\phi}_j \leftarrow \tau\phi_j + (1-\tau)\overline{\phi}_j. \tag{13}$$

where τ is a weighting parameter. During training, actions are perturbed with Gaussian exploration noise $\mathcal{N}(0,\sigma_t^2)$ with exponentially decaying standard deviation $\sigma_t = \max(\sigma_{\min}, \sigma_0 \cdot \rho^t)$ to encourage exploration. At evaluation and deployment, the learned deterministic policy is used without noise to ensure consistent, reliable control.

B. Physics-Constrained Learning Objectives

To incorporate domain knowledge and improve sample efficiency, we design physics-constrained learning objectives for the actor network. Unlike traditional Physics-Informed Neural Networks (PINNs) that embed partial differential equations directly [20], our approach incorporates power system constraints through regularization terms and bounded action spaces.

During actor training, we augment the standard policy gradient objective with a physics-based regularization term that penalizes actions violating power system constraints. This regularization encourages the learned policies to respect physical limits on reactive power injections and voltage deviations, improving sample efficiency by guiding exploration toward feasible regions of the action space [7],

$$\mathcal{L}_{\text{physics}}(\theta_j) = \lambda_{\text{act}}\|\mathbf{a}_j\|_1 + \lambda_{\text{voltage}}\sum_{k\in\mathcal{Z}_j}\max(0, |V_k - 1.0| - 0.05)^2, \tag{14}$$

where λ_{act} penalizes excessive control effort, and λ_{voltage} penalizes predicted voltage deviations beyond acceptable limits. The total actor loss becomes,

$$\mathcal{L}_{\text{total}}(\theta_j) = -Q_{\phi_j}(s, \mu_{\theta_1}(o_1), \ldots, \mu_{\theta_J}(o_J)) + \mathcal{L}_{\text{physics}}(\theta_j), \tag{15}$$

which is minimized during each actor update step via the augmented policy gradient with an additional term $\nabla_{\theta_j}\mathcal{L}_{\text{physics}}$ that enforces physical feasibility.

IV. Numerical Results

A. IEEE 13-Bus Test System

We instantiate the proposed multi-agent voltage control framework on the IEEE 13-bus three-phase distribution test feeder [3], a representative of real distribution systems with unbalanced loading, mixed line types, and transformers. The feeder exhibits high three-phase unbalance due to single-phase residential loads and asymmetric line configurations. We deploy $J = 3$ reactive power compensators at selected

locations across the feeder while considering phase-specific injections:

- Agent 1 (AB): $n_1 =$ Bus 7, controls ab-phase line-to-line reactive power injection Q_{AB}, with $\mathcal{Z}_1 = \{\text{Buses 1, 2}\}$.
- Agent 2 (BC): $n_2 =$ Bus 6, controls bc-phase line-to-line reactive power injection Q_{BC}, with $\mathcal{Z}_2 = \{\text{Buses 2, 3}\}$.
- Agent 3 (CA): $n_3 =$ Bus 12, controls ca-phase line-to-line reactive power injection Q_{CA}, with zone $\mathcal{Z}_3 = \{\text{Bus 3}\}$.

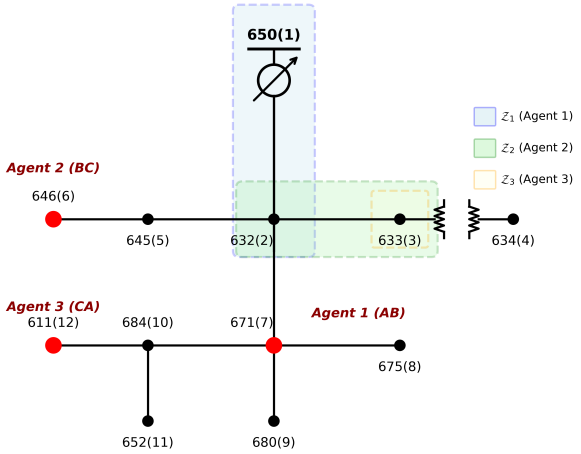

Fig. 1. IEEE 13-bus feeder with 3-agent voltage control architecture. Red circles denote the agents' reactive power control points.

Buses 2 and 3 serve as coordination points, observed by multiple agents to enable implicit coordination through overlapping zones. Each agent observes a local observation of dimension $d_s = 18$, consisting of: 15 voltage magnitudes from all monitored buses in the agent's zone and the overlapping zone (can be zero-padded), 2 dimensions of previous reactive power actions from the other two agents, and 1 dimension system feature of average VUF. The global state dimension is $J \cdot d_s = 54$. Each agent outputs a 3-dimensional action vector $\mathbf{a}_j \in \mathbb{R}^3$ representing three-phase reactive power injections, with bounds $Q_{\min} = -2$ p.u. and $Q_{\max} = 2$ p.u. The episode horizon is $T = 24$ timesteps for hourly control.

B. Experimental Configuration

1) Training and Test Data: The 3 agents were trained on 30 days of real-world data from the Baltimore-Washington region, comprising 720 hourly scenarios. Load profiles were obtained from the OpenEI Residential Load Database [21], representing TMY3-based consumption patterns for 60 aggregated homes per bus. Solar generation data were sourced from NREL's Solar Integration Studies [22], providing 5-minute resolution capacity factors resampled to hourly intervals. Training converged after 200 episodes, with the best performance achieved at episode 29 (7.125 violations per 24-hour episode).

To evaluate generalization, we reserved a held-out test set of 10 days (240 hours) from the same geographic region but temporally distinct from the training data. This test set includes diverse seasonal and load conditions not seen during training, providing a rigorous assessment of policy robustness.

2) Reward and Training Parameters: The reward weights are $w_{\text{local}} = 0.4$ and $w_{\text{global}} = 0.6$, with penalty coefficients $\lambda_{\text{local}} = 0.2$ and $\lambda_{\text{global}} = 0.1$. The tiered improvement bonuses B_{improve} provide 5, 10, 20, and 40 units for achieving 50%, 60%, 70%, and 80% improvement over baseline, respectively. The zero-violation bonus is $B_{\text{zero}} = 10$, and the VUF bonus is $B_{\text{VUF}} = 2$ for VUF below 0.5%. The coordination bonus weight is $\alpha = 0.5$. The physics regularization weights are $\lambda_{\text{act}} = 0.01$ and $\lambda_{\text{voltage}} = 0.1$.

The training parameters are specified in Table I.

TABLE I
MADDPG TRAINING HYPERPARAMETERS

Parameter	Value	Parameter	Value
Agents (J)	3	Discount (γ)	0.99
Episodes	200	Soft update (τ)	0.01
Training data	30 days	Buffer size	100k
Episode len. (T)	24 steps	Batch size (B)	64
State dim. (d_s)	18/agent	Update freq.	1
Action dim. (d_a)	3/agent	Grad. clip	≤ 1.0
Actor learning rate	10^{-4}	Init. noise (σ_0)	0.2
Critic learning rate	10^{-3}	Noise decay (ρ)	0.9995
Action bounds	$[-2, 2]$	Min. noise ($\sigma_{\min}$)	0.05

C. Control Performance

We compare against the Steinmetz control algorithm [3], a circuit-based reactive power compensation method that balances equivalent load. The baseline operates at the same 3 control points as the RL agents.

Performance is assessed using: (1) voltage violations per hour (number of bus-phase voltages outside acceptable limits), (2) VUF, (3) maximum and minimum voltages, and (4) inference time in milliseconds per action selection [23]. The test system was simulated using MATLAB, interfaced with Python via the MATLAB Engine API. The reported time includes power flow computation: Steinmetz requires iterative solutions until VUF convergence, whereas RL requires only a single forward pass plus one power flow verification. The RL agents completely eliminate voltage violations (100% reduction, from an average **13.20 to 0.00 violations per episode**) while maintaining all bus voltages within 0.999-1.002 p.u. and computational efficiency with 2.06× faster inference. The RL agents achieve zero violations across all 10 held-out test episodes, with all 12 buses maintained at 0.999-1.002 p.u. The deterministic policy (evaluated without exploration noise)

TABLE II
PERFORMANCE COMPARISON ON HELD-OUT TEST SET

Metric	Baseline	MADDPG
Violations (per episode)	13.20 ± 4.96	$\mathbf{0.00 \pm 0.00}$
Violation reduction	—	**100%**
Max voltage (p.u.)	1.000 ± 0.000	1.002 ± 0.000
Min voltage (p.u.)	0.888 ± 0.022	$\mathbf{0.999 \pm 0.000}$
VUF at Bus 2 (%)	0.00 ± 0.00	2.37 ± 0.54
Inference time (ms)	181.54	**88.21**

produces consistent perfect regulation, demonstrating effective generalization to temporally distinct scenarios not seen during training. The average VUF increases from 0.00% to 2.37%, with peak values during high-stress periods occasionally exceeding the 3% IEEE 1159 threshold. This trade-off prioritizes voltage safety—the primary operational constraint—and can be addressed in future work through constrained policy optimization to enforce per-timestep VUF limits.

We analyzed per-bus voltage regulation across the 12 buses, the baseline exhibits 13.20 violations per episode with minimum voltages as low as 0.850 p.u., demonstrating the RL agents' superior voltage regulation capability across various network conditions. The tight voltage band reflects deterministic simulation conditions; practical deployment with measurement noise may exhibit wider variation.

V. Conclusion

This paper addressed the challenge of real-time voltage control in unbalanced three-phase distribution systems by formulating the problem as a partially observable Markov game and solving it through multi-agent actor-critic deep deterministic policy gradient learning with physics-constrained objectives. The key insight is that distributed agents can achieve system-wide voltage control through implicit coordination mechanisms—overlapping zone observations and minimal inter-agent communication—without requiring centralized control during execution. The centralized training with distributed execution paradigm enables agents to learn cooperative policies while maintaining scalability. Evaluation on the IEEE 13-bus test feeder with residential load and generation data validated the approach, demonstrating perfect elimination of voltage violations while maintaining tight voltage regulation band (0.999–1.002 p.u.), fast computational speedup, and VUF below 3%. The zero-violation performance across held-out test episodes indicates that the learned policies have captured generalizable voltage-reactive power relationships rather than memorizing training patterns, addressing a critical requirement for deployment in operational distribution systems. Future work will investigate adaptive zone partitioning strategies that dynamically adjust to network topology and operating conditions, scalability to larger distribution systems and integration with active power curtailment for multi-modal control under extreme operating scenarios.

References

[1] E. Anderson, M. Ferris, A. Philpott, M. Anitescu, P. Cramton, S. Geng, R. Green, T. Homem-de Mello, O. Huber, V. Leclère *et al.*, "Ten challenges for mathematical modeling of the green-energy transition," *Current Sustainable/Renewable Energy Reports*, vol. 12, no. 1, pp. 1–13, 2025.

[2] D. K. Molzahn, L. A. Roald, J. L. Mathieu, I. A. Hiskens, D. Pinney, B. Li, K. Girigoudar, M. Yao, and S. Geng, "Mitigating phase unbalance for distribution systems with high penetrations of solar PV (final technical report)," Argonne National Laboratory (ANL), Argonne, IL (United States), Tech. Rep., 09 2021. [Online]. Available: https://www.osti.gov/biblio/1823417

[3] S. Geng and I. A. Hiskens, "Convergence of distributed Steinmetz control for balancing distribution network voltages," *IEEE Transactions on Power Systems*, vol. 37, no. 5, pp. 3370–3380, 2021.

[4] E. Dall'Anese, H. Zhu, and G. B. Giannakis, "Distributed optimal power flow for smart microgrids," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1464–1475, 2013.

[5] K. Baker, G. Hug, and X. Li, "Energy storage sizing taking into account forecast uncertainties and receding horizon operation," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 1, pp. 331–340, 2017.

[6] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE Journal of Power and Energy Systems*, vol. 4, no. 3, pp. 362–370, 2018.

[7] B. Zhang, D. Cao, W. Hu, A. M. Ghias, and Z. Chen, "Physics-informed multi-agent deep reinforcement learning enabled distributed voltage control for active distribution network using PV inverters," *International Journal of Electrical Power & Energy Systems*, vol. 155, p. 109641, 2024.

[8] D. Cao, W. Hu, J. Zhao, G. Zhang, B. Zhang, Z. Liu, Z. Chen, and F. Blaabjerg, "Reinforcement learning and its applications in modern power and energy systems: A review," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1029–1042, 2020.

[9] 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, 2020.

[10] Y. Zhang, J. Wang, and Z. Li, "Deep reinforcement learning for power system applications: An overview," *CSEE Journal of Power and Energy Systems*, vol. 6, no. 1, pp. 213–225, 2020.

[11] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multi-agent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156–172, 2008.

[12] Y. Ji, J. Wang, J. Xu, X. Fang, and H. Zhang, "Real-time energy management of a microgrid using deep reinforcement learning," *Energies*, vol. 12, no. 12, p. 2291, 2019.

[13] R. Lu, S. H. Hong, and M. Yu, "Demand response for home energy management using reinforcement learning and artificial neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6629–6639, 2019.

[14] Y. Chen, Y. Xu, Z. Li, and X. Feng, "Optimally coordinated dispatch of combined-heat-and-electrical network with demand response," *IET Generation, Transmission & Distribution*, vol. 13, no. 11, pp. 2216–2225, 2019.

[15] W. Pinthurat and B. Hredzak, "Multi-agent deep reinforcement learning for mitigation of unbalanced active powers using distributed batteries in low voltage residential distribution system," *Electric Power Systems Research*, vol. 245, p. 111599, 2025.

[16] S. Keren, C. Essayeh, S. V. Albrecht, and T. Morstyn, "Multi-agent reinforcement learning for energy networks: Computational challenges, progress and open problems," 2024.

[17] A. Bernstein, C. Wang, E. Dall'Anese, J.-Y. Le Boudec, and C. Zhao, "Load flow in multiphase distribution networks: Existence, uniqueness, non-singularity and linear models," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 5832–5843, 2018.

[18] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," 2020. [Online]. Available: https://arxiv.org/abs/1706.02275

[19] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[20] M. Raissi, P. Perdikaris, and G. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[21] National Renewable Energy Laboratory, "OpenEI: Open Energy Information - Commercial and Residential Hourly Load Profiles," https://openei.org/doe-opendata/dataset/commercial-and-residential-hourly-load-profiles, 2013, accessed: 2024.

[22] G. Brinkman, P. Denholm, E. Drury, R. Margolis, and M. Mowers, "Toward a solar-powered grid," National Renewable Energy Laboratory, Golden, CO, Tech. Rep. NREL/TP-6A20-64746, 2016.

[23] G. Guo, M. Zhang, Y. Gong, and Q. Xu, "Safe multi-agent deep reinforcement learning for real-time decentralized control of inverter based renewable energy resources considering communication delay," *Applied Energy*, vol. 349, p. 121648, 2023.