Physics Guided Reinforcement Learning for Voltage-Aware EV Charging Coordination with PV and V2G Interaction

Taniya Manzoor
Dept. of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi, India
eez228195@ee.iitd.ac.in

Brejesh Lall
Dept. of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi, India
brejesh@ee.iitd.ac.in

Bijaya Ketan Panigrahi
Dept. of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi, India
bkpanigrahi@ee.iitd.ac.in

***Abstract*—This work investigates the grid-coupled dynamics of distributed photovoltaic (PV) generation and bidirectional Vehicle to Grid (V2G) operation within a reinforcement learning based electric vehicle (EV) charging control. Building on the physics informed Twin Delayed Deep Deterministic Policy Gradient (PI TD3) framework, two controlled configurations are examined, one excluding photovoltaic generation and another restricting Vehicle to Grid (V2G) discharge, to isolate the individual effects of renewable support and bidirectional energy exchange. The framework embeds differentiable power flow and battery dynamics within the learning process, enabling physics consistent policy updates. Experiments on a modified IEEE 34 bus network with 150 EV chargers show that removing PV increases operational cost by 6% and voltage violations by 23%, while disabling V2G leads to a 12% cost rise and 17% higher voltage deviation. These results confirm that renewable generation and bidirectional flexibility can be crucial for achieving stable, efficient, and user centric grid operation, demonstrating the potential of physics guided reinforcement learning for scalable and resilient grid control.**

***Index Terms*—Physics-informed reinforcement learning, Electric Vehicle (EV) Integration, Vehicle to Grid (V2G), voltage regulation, smart grid control,**

I. Introduction

As electricity grids evolve from centralized to distributed architectures, ensuring voltage stability has become a growing concern for network operators. Electric vehicles (EVs) and distributed renewable energy sources (DERs), often viewed as challenges for grid stability, can together provide unprecedented flexibility for voltage regulation [1], [2]. The bidirectional power exchange enabled by vehicle-to-grid (V2G) technology transforms electric vehicles from passive consumers into active participants in grid operation. Coordinating V2G operations with renewable generation establishes a flexible control mechanism that counteracts renewable intermittency and supports distribution network stability [3]. This synergy enhances the effective utilization of renewable energy resources, supports voltage stability, and reduces the need for additional grid reinforcement [4].

Preliminary studies on voltage regulation using distributed energy resources and EVs relied on heuristic and optimization-based controls, such as MPC [5] and stochastic programming [6], which improved constraint handling but suffered from high computational cost and forecast dependence [7]. Motivated by demand-side flexibility, price-based and aggregator models later introduced demand-side coordination, yet offered limited direct control over voltage profiles [8].

To overcome the dual challenges of computational scalability and model dependency, reinforcement learning (RL) has evolved as a promising paradigm capable of learning real-time decision-making policies directly through interaction with stochastic, high-dimensional grid environments [9]. Hierarchical control frameworks have adopted model-free discrete approaches, including Deep Q-Networks (DQN), to achieve joint optimization of EV charging and Volt–VAR management, demonstrating improved voltage stability and energy utilization [10]. Continuous-action reinforcement learning algorithms such as Deep Deterministic Policy Gradient (DDPG) [11] and Twin Delayed DDPG (TD3) have facilitated fine-grained coordination of EV fleet charging while accounting for network voltage limitations and battery degradation effects arising from V2G operations [12]. Beyond EV-specific coordination, other actor–critic frameworks such as Proximal Policy Optimization (PPO) and Soft Actor–Critic (SAC) have demonstrated effectiveness in grid applications, including Volt–VAR optimization and remedial action schemes, demonstrating strong adaptability and stability in large-scale power systems [13], [14]. To ensure operational safety, Constrained RL and Safe RL paradigms formally incorporate power flow and voltage constraints within the policy optimization objective, guaranteeing reliable grid interaction [15]. Model-based RL enhances sample efficiency and policy robustness by leveraging learned models of network dynamics for predictive planning [16]. For scalability, Multi-agent RL (MARL) distributes the control problem, enabling decentralized voltage regulation via localized agent interactions allowing to perform online coordinated control [17]. Despite their promise, RL methods remain fundamentally challenged by the sample inefficiency of exploring high-dimensional state-action spaces, which critically undermines their reliability and constraint satisfaction

979-8-3315-8138-1

in large-scale, safety-critical grid systems [18].

Consequently, recent research has introduced Physics-Informed Reinforcement Learning (PI-RL) to embed domain knowledge directly within the training framework, moving beyond weak couplings such as penalty-based constraint handling [19]. In these methods, physical equations such as power flow relationships and battery state-of-charge dynamics are incorporated into the learning objective itself, as demonstrated by approaches like physics-informed Twin Delayed DDPG (PI-TD3) developed for large-scale coordination of EV smart charging under explicitly modeled distribution voltage constraints. [20]. This deep integration through differentiable physical models provides richer gradient information, enabling faster convergence, improved constraint satisfaction, and enhanced voltage regulation in large-scale EV coordination tasks.

Building on the PI-TD3 algorithmic framework, this work leverages the existing method for a systematic physics-driven evaluation of PV and V2G effects in grid-aware EV charging, using controlled variants in EV2Gym [21]. Our primary contributions are as follows:

- An analysis quantifying how photovoltaic generation and bidirectional EV charging influence key performance indicators, including voltage stability, operational cost, and user satisfaction.
- A systematic evaluation determining the distinct operational behaviors of CAFAP, PI-TD3, and MPC under photovoltaic and V2G-enabled settings.
- Controlled experiments demonstrating that excluding PV or V2G interactions leads to measurable degradation in voltage regulation and cost efficiency, highlighting the critical role of these physical couplings in grid performance.

II. Physics-Informed RL Modeling for EV Charging

To capture the sequential decision-making nature of coordinated EV charging, the problem is formulated as a Markov Decision Process (MDP) described by $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$. Here, $\mathcal{S}$ represents the state variables, $\mathcal{A}$ the set of possible actions, $\mathcal{P}$ the transition probabilities, $\mathcal{R}$ the reward structure, and $\gamma \in (0, 1]$ the temporal discount factor. The system's operational status at time t is represented by the state vector:

$$\mathbf{s}_t = \{\sin(\omega t),\ \cos(\omega t),\ \lambda_t,\ \mathbf{P}_t,\ \mathbf{Q}_t,\ \mathbf{\Sigma}_t,\ \mathbf{\Delta}_t,\ \mathbf{L}_t\}, \quad (1)$$

where:

- $\sin(\omega t)$ and $\cos(\omega t)$ represent the projected time-of-day using a cyclic encoding.
- λ_t denotes the prevailing real-time electricity price.
- $\mathbf{P}_t = [P_{1,t}, \ldots, P_{N,t}]$ and $\mathbf{Q}_t = [Q_{1,t}, \ldots, Q_{N,t}]$ are the vectors of net active and reactive power injections at each bus $n \in \mathcal{N}$. The net active power injection $P_{n,t}$ is defined as $P^{\mathrm{PV}}_{n,t} - P^{\mathrm{L}}_{n,t}$, accounting for local photovoltaic generation and load consumption.
- $\mathbf{\Sigma}_t = [\Sigma_{1,t}, \ldots, \Sigma_{|I|,t}]$ represents the state-of-charge of each connected electric vehicle (EV).
- $\mathbf{\Delta}_t = [\Delta_{1,t}, \ldots, \Delta_{|I|,t}]$ indicates the remaining duration until each EV's scheduled departure.
- $\mathbf{L}_t = [L_1, \ldots, L_{|I|}]$ is the lookup vector that maps each EV charger to its corresponding network bus index.

The control action executed by the agent at each decision interval t is defined by the vector:

$$a_t = [a_{1,t}, \ldots, a_{|I|,t}]^\top \in [-1, 1]^{|I|}, \quad (2)$$

where $a_{i,t} > 0$ denotes normalized charging for charger i, $a_{i,t} < 0$ denotes normalized discharging (V2G), and $a_{i,t} = 0$ indicates no action. Within the MDP framework, the transition function $\mathcal{P}(s_{t+1} \mid s_t, a_t)$ captures the system evolution, which is determined by the applied actions, the deterministic laws of grid power flow, and the probabilistic behaviors of EV arrivals, renewable outputs, and demand profiles.

To meet the system's operational goals, the normalized, weighted reward function $\mathcal{R}(\cdot)$ is structured to jointly optimize voltage regulation, energy cost, and user satisfaction. The instantaneous reward r_t is computed as:

$$\begin{aligned} \mathcal{R}_t = \alpha \sum_{n \in \mathcal{N}} & \min(0,\ \delta - |1 - V_{n,t}|) \\ & + \beta\, \Delta t \sum_{i \in \mathcal{I}} \left(\Lambda^{\mathrm{ch}}_t P^{\mathrm{ch}}_{i,t} - \Lambda^{\mathrm{dis}}_t P^{\mathrm{dis}}_{i,t} \right) \\ & + \rho \sum_{i \in \mathcal{I}} \phi_{i,t}, \end{aligned} \quad (3)$$

where the first term imposes a penalty on deviations in voltage magnitude, where $V_{n,t}$ is the voltage at bus n and δ is a violation threshold (e.g., 0.05). The second term calculates the net cost of energy transactions, with Λ^{ch}_t and Λ^{dis}_t being the charging and discharging electricity prices, respectively. The user satisfaction term penalizes insufficient charging near departure:

$$\phi_{i,t} = \max\left(0,\ \Sigma^*_{i,t} - \Sigma_{i,t}\right) \cdot \mathbb{I}[\Delta^{\mathrm{rem}}_{i,t} < \zeta] \quad (4)$$

where $\Sigma^*_{i,t}$ is the target state-of-charge, and ζ is the time threshold for departure, and $\mathbb{I}[\cdot]$ is the indicator function. The parameters α, β, ρ balance the trade-off between voltage regulation, economic efficiency, and user satisfaction in the learned policy.

To investigate the individual influence of embedded physical processes on learning behavior and grid performance, two controlled experimental configurations are introduced. In the *PV-absent* setting, all photovoltaic injections $P^{\mathrm{PV}}_{n,t}$ are set to zero, effectively isolating the impact of renewable generation on the system's operational stability and energy cost. In the *V2G-disabled* configuration, discharging actions are constrained such that $P^{\mathrm{dis}}_{i,t} = 0$, removing bidirectional energy exchange between EVs and the grid. These controlled ablations are critical to identify the specific contributions of renewable support and vehicle-to-grid flexibility within the physics-guided reinforcement learning framework, thereby providing deeper insight into how physical couplings shape voltage regulation, and economic outcomes.

979-8-3315-8138-1

The learning objective is to determine the optimal decision policy $\pi^*(a_t|s_t)$ capable of maximizing the long-term discounted reward expectation:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t r_t\right]. \quad (5)$$

Here, r_t represents the immediate reward obtained at time step t, γ is the discount factor that weighs future rewards relative to present ones, and $\mathbb{E}_\pi[\cdot]$ denotes the expectation under policy π. The Q-function (or action–value function) estimates the expected cumulative return obtained when the agent executes action a_t in state s_t and then follows policy π. This relationship can be recursively expressed through the Bellman equation as:

$$Q(s_t, a_t) = r_t + \gamma\, \mathbb{E}_{s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)} \left[Q(s_{t+1}, \pi(s_{t+1}))\right], \quad (6)$$

where the current estimate combines the immediate reward with the discounted expectation of future interactions, allowing iterative policy refinement based on sampled interactions.

The physics-guided TD3 framework, as summarized in Algorithm 1, embeds the underlying electrical and energy storage dynamics directly into the reinforcement learning environment. The distribution network follows a fixed-point current-injection formulation, where the complex bus voltages are iteratively updated according to

$$v_{n,t}^{(\kappa+1)} = Z_n + \sum_{n'\in\mathcal{N}} H_{nn'} \left(\frac{s_{n',t}}{v_{n',t}^{(\kappa)}}\right), \quad (7)$$

where Z_n represents the slack-bus voltage reference, $H_{nn'}$ is the network coupling coefficient derived from the inverse of the bus admittance matrix, and $s_{n,t}$ denotes the complex power injection at bus n, expressed as

$$s_{n,t} = (P_{n,t}^{\mathrm{PV}} - P_{n,t}^{\mathrm{L}} + P_{n,t}^{\mathrm{EV}}) + j(Q_{n,t}^{\mathrm{PV}} - Q_{n,t}^{\mathrm{L}}). \quad (8)$$

After K iterations, the converged nodal voltages are obtained as $V_{n,t} = v_{n,t}^{(K)}$, ensuring physically consistent grid responses to control actions and load variations. The energy dynamics of each EV battery adhere to bounded conservation principles, expressed as

$$\Sigma_{i,t+1} = \begin{cases} 1, & \text{if } \chi_{i,t+1} > 1, \\ \Sigma_{i,t}, & \text{if } \chi_{i,t+1} < \Sigma_{i,t}, \\ \chi_{i,t+1}, & \text{otherwise.} \end{cases} \quad (9)$$

$$\chi_{i,t+1} = \Sigma_{i,t} + \frac{\Delta t\, a_{i,t} P_i}{E_i^{\max}}. \quad (10)$$

This formulation enforces state-of-charge limits and energy balance, defining physics-consistent state transitions and reward evaluation (3), as shown in Algorithm 1, in which actor updates use a differentiable K-step rollout through the power flow equations (7)–(8) and battery dynamics (9)–(10), enabling direct gradient backpropagation to the policy network.

Algorithm 1 Physics-Informed Twin Delayed Deep Deterministic Policy Gradient (PI-TD3)

1: Initialize actor π_θ, critics Q_{ψ_1}, Q_{ψ_2}, target networks $\pi_{\theta'}, Q_{\psi_1'}, Q_{\psi_2'}$ with identical parameters and replay buffer $\mathcal{D}$
2: **for** $t = 1$ to T **do**
3: Select action: $\mathbf{a}_t \leftarrow \pi_\theta(\mathbf{s}_t) + \varepsilon,\ \ \varepsilon \sim \mathcal{N}(0, \sigma)$
4: Implement $\mathbf{a}_t$, update with $(r_t, \mathbf{s}_{t+1})$, and append $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ **to** $\mathcal{D}$
5: Access replay memory $\mathcal{D}$ to gather B transition samples $\{(\mathbf{s}_i, \mathbf{a}_i, r_i, \mathbf{s}_i')\}$
6: Add target noise: $\tilde{\mathbf{a}}_i' \leftarrow \pi_{\theta'}(\mathbf{s}_i') + \mathrm{clip}(\mathcal{N}(0,\sigma), -c, c)$
7: Compute targets: $y_i = r_i + \gamma \min_j Q_{\psi_j'}(\mathbf{s}_i', \tilde{\mathbf{a}}_i')$
8: Update critics: $\psi_j \leftarrow \psi_j - \eta_Q \nabla_{\psi_j} \frac{1}{B}\sum_i (Q_{\psi_j}(\mathbf{s}_i, \mathbf{a}_i) - y_i)^2$
9: Physics-guided actor update: Compute physics-based reward objective $J_\pi = -\sum_{k=0}^{K-1} \gamma^k R_\phi(\mathbf{s}_k, \pi_\theta(\mathbf{s}_k))$
10: where $\mathbf{s}_{k+1} = T_\phi(\mathbf{s}_k, \pi_\theta(\mathbf{s}_k))$ via K-step rollout
11: Update actor: $\theta \leftarrow \theta + \eta_\pi \nabla_\theta J_\pi$ ▷ Gradients flow through R_ϕ, T_ϕ
12: Soft-update targets: $\psi_j' \leftarrow \tau\psi_j + (1-\tau)\psi_j',\ \ \theta' \leftarrow \tau\theta + (1-\tau)\theta'$
13: **end for**

III. Experimental Results

To assess the impact of PV generation and V2G interaction, experiments were conducted using the PI-TD3 framework with two controlled variants: PV-absent (PV = 0) and V2G-disabled. The evaluation employed a modified IEEE 34-bus network with 150 charging points, using 2023 ElaadNL-based EV patterns simulated in the EV2Gym environment integrated with the RL-ADN power flow module for accurate voltage estimation. Each scenario consisted of 300 steps (15-minute intervals) with voltage limits of 0.95–1.05 p.u., and objective weights $(\alpha, \beta, \rho) = (-5 \times 10^4, 1, -10)$ were used to balance the multi-objective formulation. Each EV is modeled with a 70 kWh battery, ±22 kW charging/discharging limits, stochastic arrival and departure times derived from ElaadNL data, and full V2G participation subject to connection status. All experiments were executed on a computing cluster equipped

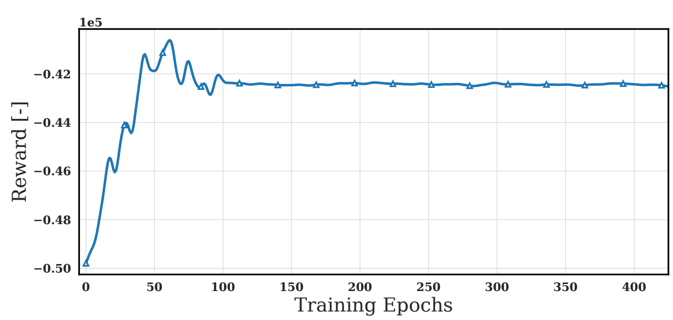

Fig. 1. Training convergence of the PI-TD3 agent showing mean reward stabilization.

979-8-3315-8138-1

TABLE I
SUMMARY OF MODEL PERFORMANCE ACROSS 25 EVALUATION SCENARIOS ON THE IEEE 34 BUS TEST SYSTEM WITH 150 CHARGERS.

Algorithm	Costs [€]	User Satisfaction [%]	Total V.V. per bus [-]	Total V.V. per step [-]	Voltage Violation [p.u.]	Tot. Energy Ch. [MWh]	Tot. Energy Dis. [MWh]
CAFAP	-2954 ±1156	100.0 ±0.0	151.04 ±132	33.6 ±21.6	-0.720 ±0.782	12.5 ±0.4	0.00 ±0.00
PI-TD3	-2105 ±362	96.4 ±2.6	97.7 ±111.4	24.2 ±27.2	-0.348 ±0.478	13.763 ±0.7	3.945 ±0.73
PI-TD3 (PV=0)	**-2238 ±1120**	**95.02 ±1.5**	**110.6 ±99.5**	**29.78 ±22.3**	**-0.412 ±0.297**	**13.615 ±0.5**	**5.023 ±0.95**
PI-TD3 (V2G-Disabled)	**-2369 ±1526**	**95.6 ±3.2**	**104.8 ±102.6**	**28.3 ±21.4**	**-0.374 ±0.472**	**13.823 ±0.8**	**0.00 ±0.00**
MPC (Oracle)	-1962 ±0.245	99.6 ±0.4	82.44 ±147.0	21.08 ±24.9	-0.314 ±0.166	26.44 ±1.5	13.84 ±0.67

with six NVIDIA GTX 1080 GPUs. At the studied scale, PI-TD3 inference runs in ∼10 ms per step, whereas MPC optimization can take several minutes.

Figure 1 illustrates the convergence behavior of the agent's mean reward over training epochs. The reward increases steadily during the initial learning phase (up to approximately 50 epochs), followed by smooth stabilization around a steady value of -0.43×10^5, indicating successful policy convergence. The rapid rise in early epochs reflects effective exploration and policy improvement, while the subsequent plateau demonstrates training stability and consistent performance across episodes.

Table I summarizes the average performance across 25 test scenarios, corresponding to independent stochastic realizations of load, PV, and EV behavior from different simulation dates and random seeds, providing implicit robustness to this variability. Relative to PI-TD3, the PV-absent case increased cost by 6% and voltage violations by 23%, while the V2G-disabled variant showed a 12% cost rise and 17% higher voltage deviation. The PV-absent setup also caused 27% more discharge activity, whereas the MPC (Oracle), used as an upper-bound benchmark rather than a realistic controller, arbitrages time-varying prices by exchanging more energy during low-/high-price periods to achieve lower net cost. Both cases show reduced user satisfaction, indicating that removing these physical couplings degrades grid stability, cost efficiency, and charging reliability.

Figure 2 illustrates the voltage magnitude profile at bus 24 over a representative day. The base PI-TD3 tracks the MPC reference closely, maintaining voltages within limits. The PV-absent and V2G-disabled configurations exhibit deeper and longer voltage dips, indicating reduced grid support and diminished voltage recovery capability. The CAFAP baseline experiences the largest voltage drops.

Figure 3 shows the average SoC trajectories of EVs over the evaluation horizon. The PI-TD3 and MPC achieve near-100% user satisfaction, ensuring full charging before departure. The PV-absent case shows delayed SoC recovery during peak demand, while the V2G-disabled variant remains less responsive due to missing bidirectional exchange.

Figure 4 shows the voltage magnitude distributions across all 33 buses. The base PI-TD3 maintains voltages close to the MPC reference, with similar medians and minimal violation spread. In contrast, the PV-absent and V2G-disabled cases exhibit greater deviations from the MPC lower bound, reflecting weakened voltage regulation when renewable or bidirectional flexibility is removed. Overall, both PV support and V2G participation are essential for sustaining grid-wide voltage stability within the physics-guided framework.

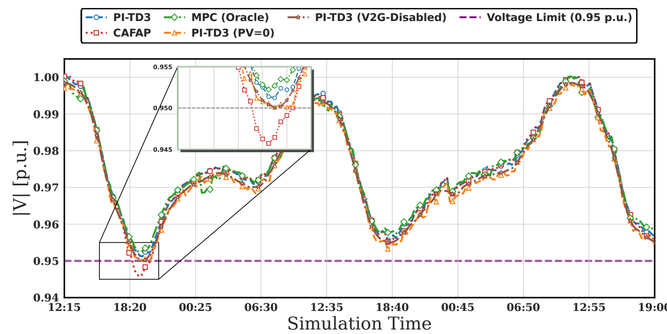

Fig. 2. Voltage magnitude profile at Bus 24.

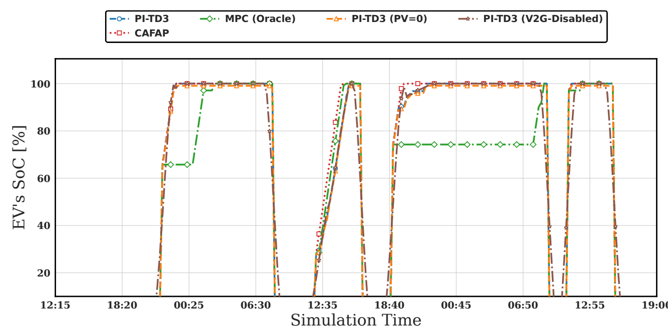

Fig. 3. SoC trajectories of EVs connected at a charger on Bus 24.

IV. CONCLUSION AND FUTURE SCOPE

This study extends the physics-guided TD3 framework to analyze the individual impacts of photovoltaic generation and bidirectional V2G interaction on coordinated EV charging and voltage regulation. Experiments on the IEEE 34-bus network reveal that excluding either renewable input or V2G flexibility increases operational costs and degrades voltage stability rel-

979-8-3315-8138-1

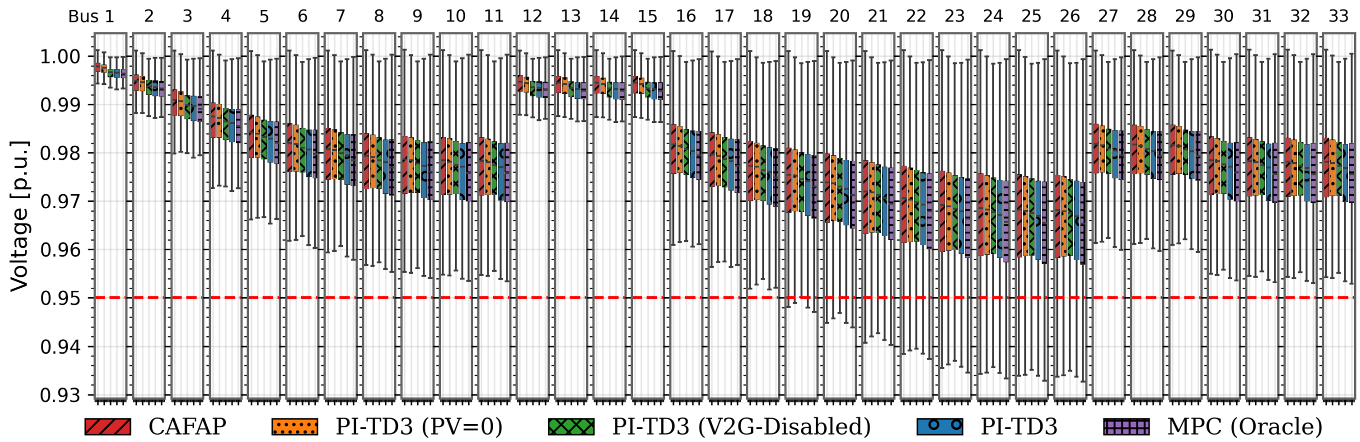

Fig. 4. Bus voltage magnitude distributions under different charging algorithms. Boxplots display median, interquartile range, and total spread, with the red dashed line denoting the 0.95,p.u. threshold.

ative to the PI-TD3 baseline and benchmark controllers. The findings highlight that renewable integration and bidirectional energy exchange are vital for efficient and stable grid operation. Looking ahead, the framework provides a foundation for intelligent energy management systems capable of adaptive, real-time decision-making in large-scale, stochastic distribution networks, bridging the gap between advanced control strategies and deployable grid-intelligent solutions, with scope for robustness and generalization studies, battery degradation, and incentive mechanisms.

References

[1] T. Su, J. Zhao, Y. Pei, Y. Yao, and F. Ding, "Analytic neural network gaussian process enabled chance-constrained voltage regulation for active distribution systems with pvs, batteries and evs," *IEEE Transactions on Power Systems*, vol. 40, no. 3, pp. 2037–2049, 2025.

[2] S. Kumar, K. R. Khan, V. L. Srinivas, G. Shankar, R. K. Saket, and K. C. Jana, "Electric vehicle fast charging integrated with hybrid renewable sources for v2g and g2v operation," *IEEE Transactions on Industry Applications*, pp. 1–14, 2025.

[3] H. N. T. Nguyen, C. Zhang, and M. A. Mahmud, "Optimal coordination of g2v and v2g to support power grids with high penetration of renewable energy," *IEEE Transactions on Transportation Electrification*, vol. 1, no. 2, pp. 188–195, 2015.

[4] S. Tirunagari, M. Gu, and L. Meegahapola, "Reaping the benefits of smart electric vehicle charging and vehicle-to-grid technologies: Regulatory, policy and technical aspects," *IEEE Access*, vol. 10, pp. 114 657–114 672, 2022.

[5] L. Wang, A. Dubey, A. H. Gebremedhin, A. K. Srivastava, and N. Schulz, "Mpc-based decentralized voltage control in power distribution systems with ev and pv coordination," *IEEE Transactions on Smart Grid*, vol. 13, no. 4, pp. 2908–2919, 2022.

[6] A. Papavasiliou and S. S. Oren, "Large-scale integration of deferrable demand and renewable energy sources," *IEEE Transactions on Power Systems*, vol. 29, no. 1, pp. 489–499, 2014.

[7] S. Vandael, B. Claessens, D. Ernst, T. Holvoet, and G. Deconinck, "Reinforcement learning of heuristic ev fleet charging in a day-ahead electricity market," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1795–1805, 2015.

[8] E. Sortomme and M. A. El-Sharkawi, "Optimal combined bidding of vehicle-to-grid ancillary services," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 70–79, 2012.

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

[10] Q. Yang, G. Wang, A. Sadeghi, G. Giannakis, and J. Sun, "Two-timescale voltage control in distribution grids using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2313–2323, 2020.

[11] D. Liu, P. Zeng, S. Cui, and C. Song, "Deep reinforcement learning for charging scheduling of electric vehicles considering distribution network voltage stability," *Sensors*, vol. 23, no. 3, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/3/1618

[12] R. Hossain, M. Gautam, M. M. Lakouraj, H. Livani, and M. Benidris, "Volt-var optimization in distribution networks using twin delayed deep reinforcement learning," in *2022 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2022, pp. 1–5.

[13] S. Song, Y. Jung, G. Jang, and S. Jung, "Proximal policy optimization through a deep reinforcement learning framework for remedial action schemes of vsc-hvdc," *International Journal of Electrical Power Energy Systems*, vol. 150, p. 109117, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0142061523001746

[14] R. Hossain, M. Gautam, M. MansourLakouraj, H. Livani, M. Benidris, and Y. Baghzouz, "Soft actor critic based volt-var co-optimization in active distribution grids," in *2022 IEEE Power Energy Society General Meeting (PESGM)*, 2022, pp. 01–05.

[15] J. Fan, A. Liebman, and H. Wang, "Safety-aware reinforcement learning for electric vehicle charging station management in distribution network," in *2024 IEEE Power Energy Society General Meeting (PESGM)*, 2024, pp. 1–5.

[16] R. R. Hossain, T. Yin, Y. Du, R. Huang, J. Tan, W. Yu, Y. Liu, and Q. Huang, "Efficient learning of voltage control strategies via model-based deep reinforcement learning," 2022. [Online]. Available: https://arxiv.org/abs/2212.02715

[17] D. Cao, W. Hu, J. Zhao, Q. Huang, Z. Chen, and F. Blaabjerg, "A multi-agent deep reinforcement learning based voltage regulation using coordinated pv inverters," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 4120–4123, 2020.

[18] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3826–3839, 2020.

[19] R. Raja Hossain, Q. Huang, K. Mahapatra, and R. Huang, *Physics-Informed Deep Reinforcement Learning-Based Control in Power Systems*, 2025, pp. 67–77.

[20] S. Orfanoudakis, F. A. Oliehoek, P. Palensky, and P. P. Vergara, "Physics-informed reinforcement learning for large-scale ev smart charging considering distribution network voltage constraints," 2025. [Online]. Available: https://arxiv.org/abs/2510.12335

[21] S. Orfanoudakis, C. Diaz-Londono, Y. Emre Yılmaz, P. Palensky, and P. P. Vergara, "Ev2gym: A flexible v2g simulator for ev smart charging research and benchmarking," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 2, pp. 2410–2421, 2025.