Safety Constrained Multi-Agent Reinforcement Learning for Active Voltage Control

Yang Qu , Jinming Ma , Feng Wu*
School of Computer Science and Technology,
University of Science and Technology of China, Hefei, China
{qu180518, jinmingm}@mail.ustc.edu.cn, wufeng02@ustc.edu.cn

Abstract

Active voltage control presents a promising avenue for relieving power congestion and enhancing voltage quality, taking advantage of the distributed controllable generators in the power network, such as roof-top photovoltaics. While Multi-Agent Reinforcement Learning (MARL) has emerged as a compelling approach to address this challenge, existing MARL approaches tend to overlook the constrained optimization nature of this problem, failing in guaranteeing safety constraints. In this paper, we formalize the active voltage control problem as a constrained Markov game and propose a safety-constrained MARL algorithm. We expand the primal-dual optimization RL method to multi-agent settings, and augment it with a novel approach of double safety estimation to learn the policy and to update the Lagrange-multiplier. In addition, we proposed different cost functions and investigated their influences on the behavior of our constrained MARL method. We evaluate our approach in the power distribution network simulation environment with real-world scale scenarios. Experimental results demonstrate the effectiveness of the proposed method compared with the state-of-the-art MARL methods.

1 Introduction

In recent years, significant progress has been witnessed in the development of renewable and distributed sources of electricity, such as rooftop photovoltaics (PVs) [Schardt and te Heesen, 2021]. While these innovations hold promise to address energy shortages and environmental concerns, they also introduce growing complexity and uncertainty into modern power systems, presenting formidable challenges. One of the notable challenges associated with the high penetration of distributed energy is the potential for voltage fluctuations exceeding the power grid standards [Till *et al.*, 2020]. Mitigating these fluctuations requires harnessing control capabilities of PV inverters and other devices. As discussed in the recent work [Wang *et al.*, 2021], achieving active voltage control across the entire network, particularly when access to global information is limited, requires intricate coordination.

Recently, Multi-Agent Reinforcement Learning (MARL) algorithms have demonstrated exceptional performance in many domains [Canese *et al.*, 2021; Wang and Wu, 2020], and MARL algorithms are gradually finding applications in real-world scenarios. In comparison to traditional voltage regulation methods like Optimal Power Flow (OPF) [Gan *et al.*, 2013; Zheng *et al.*, 2016] and droop control [Jahangiri and Aliprantis, 2013], MARL algorithms offer several advantages: 1) They follow the Centralized Training with Decentralized Execution (CTDE) schema, and correspond to the trait that usually only limited local information can be accessed in active voltage control. 2) They are data-driven and are naturally model-free, while traditional methods need exact system models. 3) Traditional methods have high computational complexity in solving power flow equations, having difficulty in real-time response. Meanwhile, MARL shows adaptability to respond to environmental changes quickly.

To date, there have been some previous efforts to apply MARL on voltage control problems and achieved promising performances [Chen *et al.*, 2022]. Previous research has explored the potential of MARL by taking various approaches, such as integrating traditional control techniques [Wang *et al.*, 2020; Sun and Qiu, 2021], enhancing state information representation and region segmentation [Cao *et al.*, 2020b], refining MARL reward designs [Wang *et al.*, 2021; Wang *et al.*, 2020] and so on, achieving remarkable advances in the field. However, the unique property of constrained optimization of this problem is less concerned in the previous work. Specifically, in the active voltage control problem, the constraint is the voltage threshold and the objective is the total power loss. Moreover, unlike most safe RL benchmarks [Ray *et al.*, 2019], the fulfillment of the constraint is the primary concern. In the power distribution network, large voltage fluctuations will affect the stability of the power system, reduce power quality for users, and even cause irreversible damage to the system equipment, resulting in system failure and collapse. It is relatively acceptable to ensure the safety of the system at the cost of higher power loss. Given the dynamic of the system, it is generally more difficult and complicated to meet the constraint than to improve the objective in this problem, for the constraints must be fulfilled given a sequence of uncontrollable events. Many real-world prob-

*Feng Wu is the corresponding author.

lems exhibit the aforementioned two features, for example autonomous driving, wireless security, traffic control [Ma and Wu, 2023] and so on. To the best of our knowledge, the application of constrained RL on these problems is less studied in the literature [Gu *et al.*, 2022].

Against this background, we propose a novel constrained MARL approach named Multi-Agent RL with Double Estimation of Lagrangian Constraint (MA-DELC). Specifically, we formulate the active voltage control problem as a constrained Markov game. Firstly, we incorporate a safety critic and a cost estimator alongside the conventional reward critic. The safety critic is used for guiding the policy to fulfill the constraint in the long term, and the cost estimator is used for updating the Lagrange-multiplier. These additional components allows our system to effectively balance the optimization objectives and the imposed constraints.

The results of our experiments highlight the efficacy of this approach, demonstrating that the incorporation of structural information significantly enhances agent performance. Additionally, we explore the conversion of voltage constraints into various cost functions, meticulously studying the impact of different cost function designs through comprehensive experimental analysis. This exploration is of great practical significance, particularly when applying constrained MARL techniques to real-world applications, where the choice of cost function can have a profound influence on algorithm behavior and performance. Our experiments on the MAPDN environment [Wang *et al.*, 2021], commonly used in the literature, show that these components enhance performance and scale well compared with the state-of-the-art RL methods.

2 Related Work

Here, we briefly review recent MARL and constrained RL approaches for active voltage control.

2.1 MARL for Active Voltage Control

Recently, MARL algorithms have gained significant attention in active voltage control. Wang *et al.* [2020] and Sun and Qiu [2021] combined traditional voltage control methods and MARL. Wang *et al.* [2020] used MADDPG [Lowe *et al.*, 2017] with a manually designed voltage inner loop, and agents set reference voltage instead of reactive power as their control actions. Sun and Qiu [2021] proposed a two-stage volt-var control method, in which the traditional optimal flow method [Gan *et al.*, 2013] is used to dispatch on-load tap changer and capacitor banks in the first stage, and in the second stage the MADDPG algorithm is used to regulate the reactive power of PVs. Gao *et al.* [2021] adopted a discrete action space of the set of all tap positions of the voltage regulating device, and proposed a consensus-based maximum entropy MARL framework composed of max-entropy MARL algorithm and a consensus strategy. Cao *et al.* [2020b] used spectral clustering to divide the power distribution network into several sub-networks according to the voltage and reactive power sensitivity, and then applied MATD3 [Ackermann *et al.*, 2019] to solve the voltage control problem, each sub-network corresponds an agent. Cao *et al.* [2020b] also used reactive power as control actions. Cao *et al.* [2021] used both reactive power and the curtailment of active power as control actions, and applied MASAC [Pu *et al.*, 2021] to solve decentralized voltage control problems with high penetration of PVs. Cao *et al.* [2020a] and Wang *et al.* [2022] combines MADDPG [Lowe *et al.*, 2017] critic with transformer to enhance inter-agent coordination.

In this paper, we adopt continuous control action space of reactive power of PV inverters, which is more flexible and reliable than discrete control action. It is worth noting that our approach does not need topological information, and maintains distributed during execution.

2.2 Constrained RL for Active Voltage Control

Now, we discuss the methods that applied constrained RL algorithms to solve the active voltage control problem.

Shi *et al.* [2022] designed a stability-constrained RL framework which utilizes a manually designed Lyapunov function, and proved stability guarantees. Wang *et al.* [2019] directly applied Constrained Policy Optimization (CPO) [Achiam *et al.*, 2017] to solve the Volt-VAR problem. Liu *et al.* [2023] utilized a safety layer to keep the action within safety limits by calculating the power flow equations based on the physical model of the power grid. The above methods are all based on single-agent settings, and are trained and executed in a centralized style.

Liu and Wu [2021] formulated the Volt-VAR Control problem as a constrained Markov game with the reactive power of inverters or static var compensators (SVCs) as control actions, and proposed a Multi-Agent Constrained Soft Actor-Critic algorithm (MACSAC) that combines MASAC [Pu *et al.*, 2021] and Lagrangian constraint. Shi *et al.* [2023] used safety layer method based on multi-agent settings, and they used neural networks to approximate cost functions, getting rid of the need of precise physical models.

Although Liu and Wu [2021] also used the Lagrangian constraint MARL method, our approach differs in the double safety estimation framework and the adaptive Lagrange multiplier. Actually, MACSAC is similar to MADELC w/o cost estimator in our ablation study in Section 5.3. Shi *et al.* [2023] focuses on the safety issue and ignores power loss reduction. Moreover, it considered only one-step cost, while our approach considers both one-step cost and the cumulative cost return, enabling agents to minimize costs more steadily.

3 Background

3.1 Active Voltage Control on Power Networks

In this paper, we consider a power distribution network installed with roof-top photovoltaics (PVs). As shown in Figure 1, we model the power distribution network as a tree graph $\mathcal{G} = (V, E)$, where $V = \{0, 1, \ldots, |V|\}$ represents the set of nodes (buses), and $E = \{1, 2, \ldots, |V|\}$ represents the set of edges (branches). Nodes in the distribution network are divided into several zones based on their shortest path from the terminal to the main branch [Gan *et al.*, 2013]. We denote regions as $\mathrm{Zone}_z, z = 1, \ldots, |Z|$. Node 0 is connected to the main grid, and has a stable voltage magnitude, denoted as v_0. Each node may be connected to load and PV. We denote the

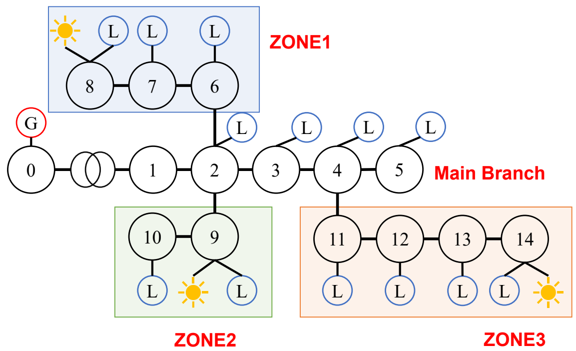


Figure 1: An example of power distribution networks. The black circles with numbers represent buses; G represents an external generator; Ls represent loads; and the sun emoji represents the location where a PV is installed. We control the bus voltage by adjusting the reactive power generation of PV inverters in buses 8, 9, and 14.

set of nodes installed with PVs in Zone$_z$ as PV$_z$, and the set of nodes connected with load in Zone$_z$ as Load$_z$.

For each bus $j \in V$, let v_j and θ_j be the magnitude and phase angle of the voltage, and $s_j = p_j + \mathbf{i}q_j$ denotes the complex power injection. Note that there are complex and nonlinear relationships between these physical quantities that satisfy the power system dynamics rules [Farivar *et al.*, 2013; Wang *et al.*, 2021]. Hence, it is hard to solve it in close form.

Let $v_0 = 1.0$ per unit ($p.u.$). For other buses, 5% voltage deviation is usually allowed for safe and optimal operation. The objective of active voltage control is to keep the nodal voltages in a normal range (e.g., 0.95 $p.u.$ to 1.05 $p.u.$) and concurrently minimize total active power loss $P_{loss} = \Sigma_{(i,j)\in E} R_{ij} I_{ij}^2$, or known as the line loss. R_{ij} is the resistance of branch (i, j) and I_{ij} is the current magnitude from bus i to j [Hu *et al.*, 2022]. The optimization problem is as follow:

$$\begin{aligned} \min \quad & P_{loss} = \textstyle\sum_{(i,j)\in E} R_{ij} I_{ij}^2 \\ \text{s.t.} \quad & 0.95p.u. \leq v_i \leq 1.05p.u., \forall i \in V\backslash\{0\}; \\ & \text{satisfy power flow dynamics.} \end{aligned} \tag{1}$$

When loads are heavy at nighttime, v_i may drop below $0.95p.u.$. Meanwhile, large power generation of PVs in the middle of the day may cause v_i rise to exceed $1.05p.u.$. These problems can be handled by applying appropriate control on the reactive power injection of the PV inverter.

3.2 Constrained Markov Game for Active Voltage Control

We consider the control process of the PV inverters as constrained Markov game [Altman and Shwartz, 1998]: $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{O}_i\}_{i\in\mathcal{N}}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \mathcal{T}, r, \Omega, \mathcal{C}, c, \gamma \rangle$, where $\mathcal{N} = \{1, 2, \dots, n\}$ is a set of n agents; $\mathcal{S}$ denotes the state space; $\mathcal{O} = \times_{i\in\mathcal{N}} \mathcal{O}_i$ denote the joint observation set, where $\mathcal{O}_i$ denotes the observation set of agent i; $\mathcal{A} = \times_{i\in\mathcal{N}} \mathcal{A}_i$ denote the joint action set, where $\mathcal{A}_i$ denotes the action set of agent i; $\mathcal{T} : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0, 1]$ is probabilistic state transition function; r is the reward function; $\Omega : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to [0, 1]$ denotes the perturbation of the observers for agents' joint observations over the states after decisions; $\mathcal{C} : \mathcal{S} \times \mathcal{A} \to \mathcal{R}$ is the cost function; c denote the cost limit. MARL algorithms for constrained Markov game aim to search a policy π and solve this constrained optimization problem:

$$\begin{aligned} \max\nolimits_\pi \quad & \mathbb{E}_{(s_t,a_t)\sim\pi} \left[\textstyle\sum_t \gamma^t r(s_t, a_t)\right], \\ \text{s.t.} \quad & \mathbb{E}_{(s_t,a_t)\sim\pi} \left[\textstyle\sum_t \gamma^t \mathcal{C}_i(s_t, a_t)\right] < c_i, i = 1, 2, \dots, m. \end{aligned} \tag{2}$$

In the context of the active voltage control problem within the Constrained Markov Game framework, the essential elements are defined as follows:

- **Agent**: Each agent is responsible for controlling an individual PV inverter in the system.
- **Observation**: The observation of an agent is composed of relative physical quantities of the buses in the region to which the agent belongs. Specifically, $\mathcal{O}_i = \mathcal{L}_r \times \mathcal{P}_r \times \mathcal{V}_r$, where $\mathcal{L}_r = \{p_j^L, q_j^L > 0 | \forall V_j \in \text{Load}_r\}$: p_j^L, q_j^L are active and reactive power of the load connected to node j respectively; $\mathcal{P}_r = \{p_j^{PV}, q_j^{PV} > 0 | \forall V_j \in \text{PV}_r\}$: p_j^{PV}, q_j^{PV} are active and reactive powers generated by PV inverters in node j; $\mathcal{V}_r = \{v_j > 0, \theta_j \in [-\pi, \pi] | \forall V_j \in \text{Zone}_z\}$, v_j, θ_j is voltage magnitude and phase of node j. Agents in the same region share the same observation.
- **Action**: Each agent $i \in \mathcal{N}$ has a continuous action set $\mathcal{A}_i = \{a_i : -1.0 \leq a_i \leq 1.0\}$ that denotes the ratio of maximum reactive power it can generate.
- **State and State Transition**: The global state is composed of bus voltage, the active and reactive power of loads and PVs of all nodes. Power of loads and active power generation of PVs of the next state are obtained from the dataset, and other parts are calculated by solving power dynamic equations with data from the dataset, current state and action.
- **Reward**: Agents receive reward at each time step: $r = -l_q(q^{PV}) = -\frac{1}{|I|}||q^{PV}||$, which is the mean reactive power generation loss of PVs, and is an approximation of power loss caused by PVs.
- **Constraint**: According to Eq.1, the voltage safety constraint corresponds to the cost function $\mathcal{C}(s, a) = \mathbb{I}((v_i < 0.95||v_i > 1.05)|\exists i \in V\backslash\{0\})$. We also propose other cost functions in order to provide more information and guide the agents, which are described in section 4.2. During training, the algorithm utilizes normalized rewards and costs, with cost values normalized to the range (-1,1). We expect that all bus voltages remain within the safe range, and the cost limit c is set to -0.5 under the normalized standard.

4 Method

We propose the Multi-Agent reinforcement learning with Double Estimation of Lagrangian Constraint (MA-DELC) algorithm and study the impact of different cost function designs. These components are introduced below in details.

4.1 MARL with Double Estimation of Lagrangian Constraint

Constrained optimization problems in Eq. 2 can be solved by the Lagrange-multiplier method [Bertsekas, 2014]. Specifically, it introduces a Lagrangian function $\mathcal{L}(\pi, \alpha)$ as defined:

$$\begin{aligned} &\mathcal{L}(\pi, \alpha) \doteq f(\pi) - \alpha(g(\pi) - c), \text{where} \\ &f(\pi) = \mathbb{E}_{(s_t, a_t) \sim \pi}\left[\sum_t \gamma^t r(s_t, a_t)\right], \text{and} \\ &g(\pi) = \mathbb{E}_{(s_t, a_t) \sim \pi}\left[\sum_t \gamma^t \mathcal{C}(s_t, a_t)\right]. \end{aligned} \tag{3}$$

where $\alpha > 0$ is the Lagrange-multiplier or the dual variable; $f(\pi)$ is the discounted cumulative return of reward; $g(\pi)$ is the discounted cumulative return of cost; and c is the cost limit. Here, $\mathcal{L}(\pi, \alpha)$ can be viewed as the optimization objective, and coefficient α determines the degree of emphasis on the constraint $g(\pi)$.

In MA-DELC, two separate critics (Q functions) are trained, where one is a reward-critic Q_ϕ^r for estimating the cumulative reward return $f(\pi)$, and the other one is a cost-critic Q_ϕ^c for estimating $g(\pi)$, ϕ denote the parameters of the critic networks. Follow the formulation of SAC-Lagrange [Ha *et al.*, 2020] and MADDPG [Lowe *et al.*, 2017], we use TD-error loss to train critics:

$$\begin{aligned} J_Q(\phi) &= \mathbb{E}_{(s_t, a_t) \sim \mathcal{D}}\left[(Q_\phi^r(s_t, a_t) - y^r)^2 + (Q_\phi^c(s_t, a_t) - y^c)^2\right], \\ y^r &= r_t + \gamma \bar{Q}_{\phi'}^r(s_{t+1}, a_{t+1})|_{(r_t, s_{t+1}) \sim \mathcal{D}, a_{t+1} \sim \pi_{\theta'}(s_{t+1})}, \\ y^c &= c_t + \gamma \bar{Q}_{\phi'}^c(s_{t+1}, a_{t+1})|_{(c_t, s_{t+1}) \sim \mathcal{D}, a_{t+1} \sim \pi_{\theta'}(s_{t+1})}. \end{aligned} \tag{4}$$

$\bar{Q}_{\phi'}^r, \bar{Q}_{\phi'}^c$ denotes the target critic network with the parameter ϕ' that is soft copied from the critic network every τ steps.

In existing primal-dual constrained RL methods [Ha *et al.*, 2020; Liang *et al.*, 2018], the cost-critic Q^c is directly used for updating the dual variable, which is inaccurate due to the uncertainty of cumulative cost and the bias of the Q value approximation. This may lead to the algorithm overestimating the constraint satisfaction. To avoid these errors, we train a one-step cost estimator $\hat{C}$ with the parameter ψ for adjusting the dual variable α with the mean square error.

$$J_{\hat{C}}(\psi) = ||\hat{C}(s_t, a_t) - c_t||_2^2 | (s_t, c_t) \sim \mathcal{D}, a_t \sim \pi_\theta(s_t) \tag{5}$$

On the purpose of maximizing the Lagrangian function $\mathcal{L}(\pi, \alpha)$, we can get the actor loss:

$$J_\pi(\theta) = \mathbb{E}_{s_t \sim \mathcal{D}, a_t \sim \pi_\theta}\left[-Q_\phi^r(s_t, a_t) + \alpha Q_\phi^c(s_t, a_t)\right] \tag{6}$$

where $\mathcal{D}$ is the replay buffer and θ is the parameters of the actor network.

To ensure that the constraint $g(\pi) < c$ is satisfied, we learn the adaptive safety weight α (Lagrangian multiplier) by minimizing the loss $J(\alpha)$:

$$J(\alpha) = \mathbb{E}_{s_t \sim \mathcal{D}, a_t \sim \pi_\theta}\left[\alpha(c - \hat{C}_\psi(s_t, a_t))\right] \tag{7}$$

In this case, if $\hat{C}_\psi(s_t, a_t) > c$, the constraint is not satisfied, so α will be increased to emphasize safety more; otherwise α will be decreased.

Algorithm 1 Training process of MA-DELC

```
Initialize replay buffer D, actor and critic parameters.
for each episode do
    for each time step t do
        for agent i ∈ N do
            Select action a_i = μ_θi(o_i) + ξ, ξ ~ N(0, Σ_std).
        end for
        Execute joint action a = (a_1, ..., a_n) at state s.
        Get reward r, cost c by going to next state s'.
        D ← D ∪ (s, a, r, c, s'), s ← s'.
        Sample batch B from D.
        Update Q^r_φ, Q^c_φ by minimizing the loss in Eq. 4.
        Update Ĉ_ψ by minimizing the loss in Eq. 5.
        Update π_θ by minimizing the loss in Eq. 6.
        Update α by minimizing the loss in Eq. 7.
        θ' ← τθ' + (1 − τ)θ, φ' ← τφ' + (1 − τ)φ .
    end for
end for
```

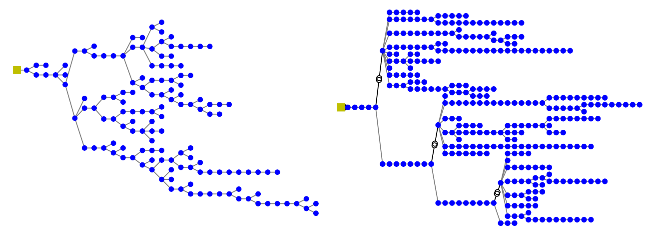


Figure 2: The topology of the distribution network of the 141-bus and 322-bus scenario in the MAPDN environment.

4.2 Cost Functions

In Section 3, we formalize the cost function according to the form of the constraint and the Lagrange-multiplier method. However, this boolean cost function is not informative enough for agents to learn from in large-scale scenarios. To provide more information about the constraint and the degree of violation, we propose several cost functions below:

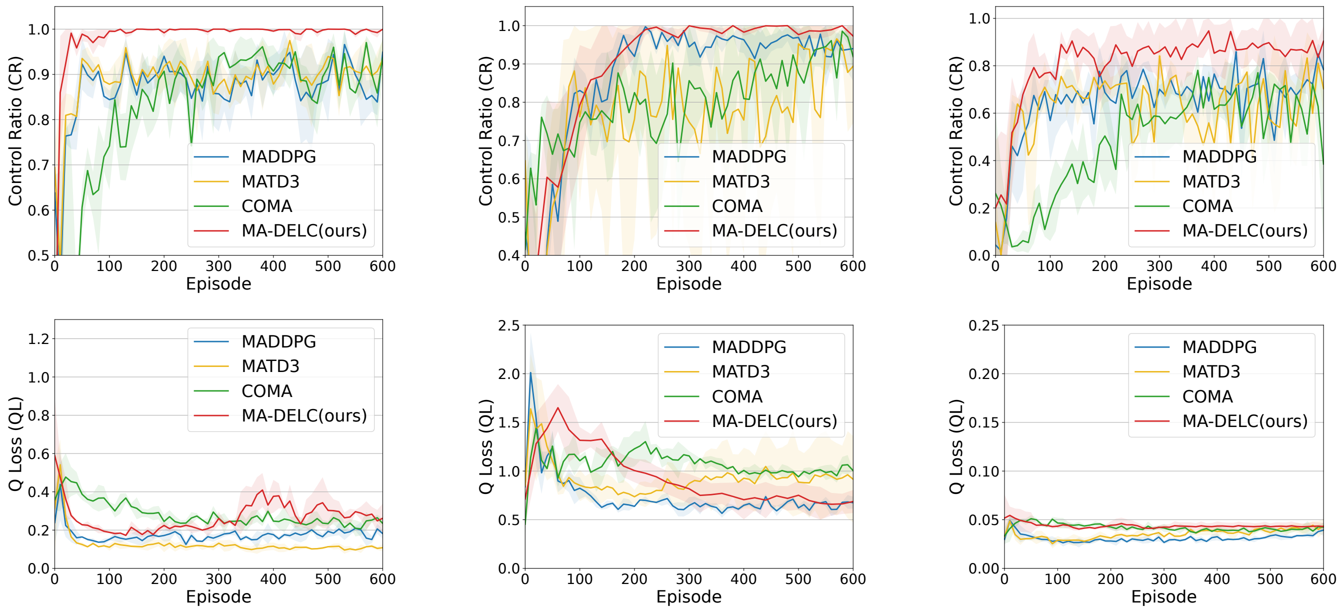
1. **Boolean Cost**: This is the basic cost function that implies whether the constraint is satisfied. If all buses' voltages are in the safe range at the time step, the cost is 0; otherwise, the cost is 1. Denote

$$C_{percent} = 1 - \frac{\sum_i \mathbb{I}(v_i < 0.95 || v_i > 1.05)}{|V|} \tag{8}$$

as the percent of buses whose voltages are within the safety limit. Then this boolean cost function is as:

$$\mathcal{C}(s, a) = \mathbb{I}(C_{percent} = 1.0). \tag{9}$$

2. **Step Cost**: We construct the piece-wise cost function to provide more information about the whole picture of the power grid. This cost function differs from the boolean cost function in that it returns a lower cost when more than 90% of all buses are under control. We expect the step cost function to help the agents to distinguish between safe and unsafe actions when the constraint is hard



(a) 33-bus scenario (b) 141-bus scenario (c) 322-bus scenario

Figure 3: Control Ratio (CR) and Q Loss (QL) results of the different algorithms (CR: higher is better, QL: lower is better).

to be fulfilled completely and the agents cannot get information from the boolean cost function.

$$\mathcal{C}(s,a) = \begin{cases} 0, & C_{percent} = 1.0 \\ 0.5, & C_{percent} \in [0.9, 1.0) \\ 1, & C_{percent} \in [0, 0.9) \end{cases} \tag{10}$$

3. **V-Loss Cost**: The v-loss cost function measures the mean voltage deviation from $1.0p.u.$ of all buses. This cost function is continuous and is the most informative one among other cost functions. Minimizing v loss means keep nodal voltage as stable as possible, and correlates with safety. However, it does not give explicit information about whether the constraint is violated.

$$\mathcal{C}(s,a) = \frac{1}{|V|}\sum_{i \in V} l_v(v_i) = \frac{1}{|V|}\sum_{i \in V}(|v_i - 1.0|_1) \tag{11}$$

4.3 Implementation Details

The main procedures of our MA-DELC method are outlined in Algorithm 1. In lines 4-9, agents interact with the environment to obtain transition data. We add noise ξ on the action to be executed to encourage exploration. The variance Σ_{std} of the noise ξ is a hyper-parameter. Actors of different agents use shared parameters, and the input of actors is local observation o_i appended with agent id. Next, we update α and the parameters of actor, critic, cost estimator and target critic networks in lines 10-15.

5 Experiments

5.1 Experiment Setups

Environment. We conduct experiments based on the Multi-Agent Power Distribution Networks (MAPDN) environment [Wang *et al.*, 2021]. The MAPDN environment serves as a suitable framework for distributed active voltage control, readily accommodating the application of MARL algorithms. This environment presents 3 scenarios: the 33-bus, the 141-bus and the 322-bus scenarios, containing 6, 22 and 38 agents, 32, 84 and 337 loads, respectively. The load and PV data is sourced from real-world datasets, lending authenticity to the simulations. The topological structures of the 141-bus and 322-bus networks are visualized in Figure 2.

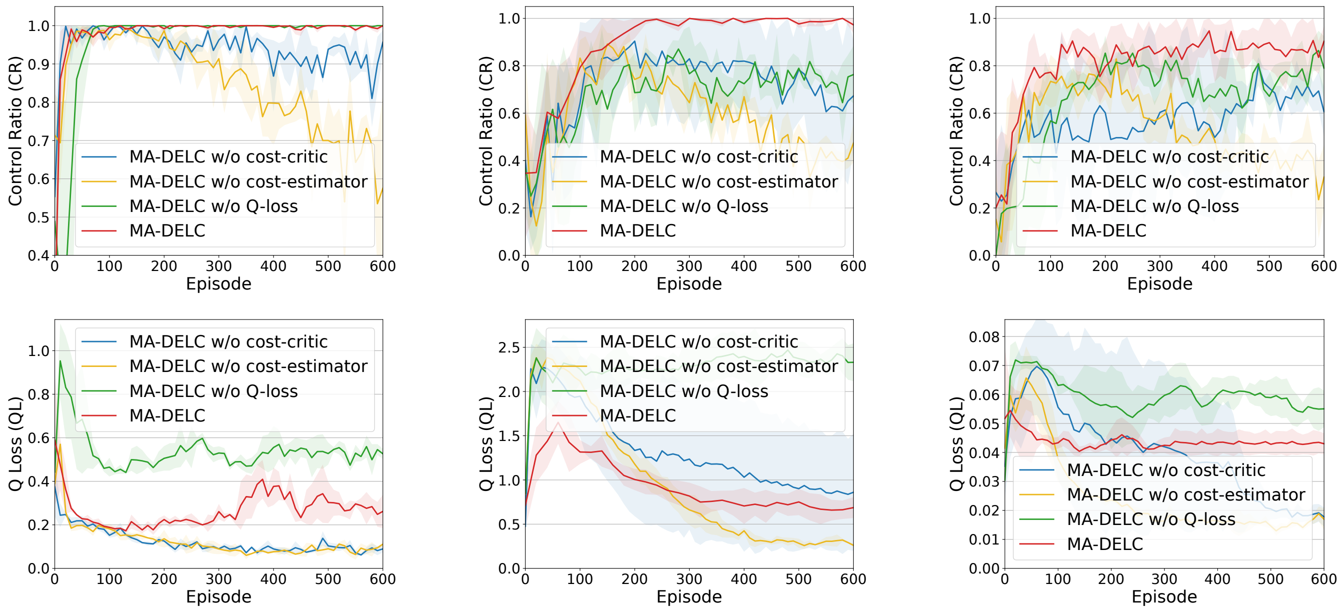
To assess the performance of the MARL algorithms in these scenarios, we employ two key evaluation metrics:

- **Controllable Ratio (CR)**: It calculates the ratio of the steps in an episode where all buses' voltage is under control within the safety range (i.e. $0.95 \le v_i \le 1.05, \forall i \in V\backslash\{0\}$).
- **Q Loss (QL)**: It calculates the mean reactive power generations by agents per time step. This metric is chosen as a proxy for power loss (PL), which is often challenging to obtain for the entire network. QL serves as an approximation of the power loss attributable to PVs (agents).

These metrics together enable a comprehensive assessment of the algorithms' capabilities in terms of voltage control and power loss. In our evaluation, our focus is to identify algorithms that achieve high CR while maintaining low QL, with CR being the primary concern.

Baseline Methods. According to the experimental results of [Wang *et al.*, 2021], COMA [Foerster *et al.*, 2017], MADDPG [Lowe *et al.*, 2017] and MATD3 [Ackermann *et al.*, 2019] surpass other MARL algorithms in the MAPDN environment. Therefore, we use them as our baselines. We provide brief descriptions of each of these baselines below:

- **COMA** [Foerster *et al.*, 2017]: the method incorporates



(a) 33-bus scenario (b) 141-bus scenario (c) 322-bus scenario

Figure 4: Control Ratio (CR) and Q Loss (QL) results of the ablation studies (CR: higher is better, QL: lower is better).

counterfactual reasoning to address credit assignment problems in cooperative multi-agent settings.

- **MADDPG** [Lowe *et al.*, 2017]: the method employs actor-critic architectures and centralized training with decentralized execution to enhance the coordination of agents. In the architecture, each agent has a centralized critic and decentralized actor.
- **MATD3** [Ackermann *et al.*, 2019]: the method combines the advantages of actor-critic methods with twin critics and target policy smoothing, making it suitable for complex multi-agent environments.

The reward of these non-Lagrangian baselines follows the settings in [Wang *et al.*, 2021], defined as $r = -\frac{1}{|V|}\sum_{i\in V} l_v(v_i) - \beta \cdot l_q(q^{PV})$. $l_v(\cdot)$ is a voltage barrier function to measure voltage deviation from $1.0p.u.$; β is a hyper-parameter to trade-off between satisfying safety constraints and reducing power loss; $l_q(q^{PV})$ is the reactive power generation loss of PVs.

5.2 Results

The median results of CR and QL of various algorithms are shown in Figure 3, and the shaded areas in the charts represent the standard deviations of runs with different random seeds.

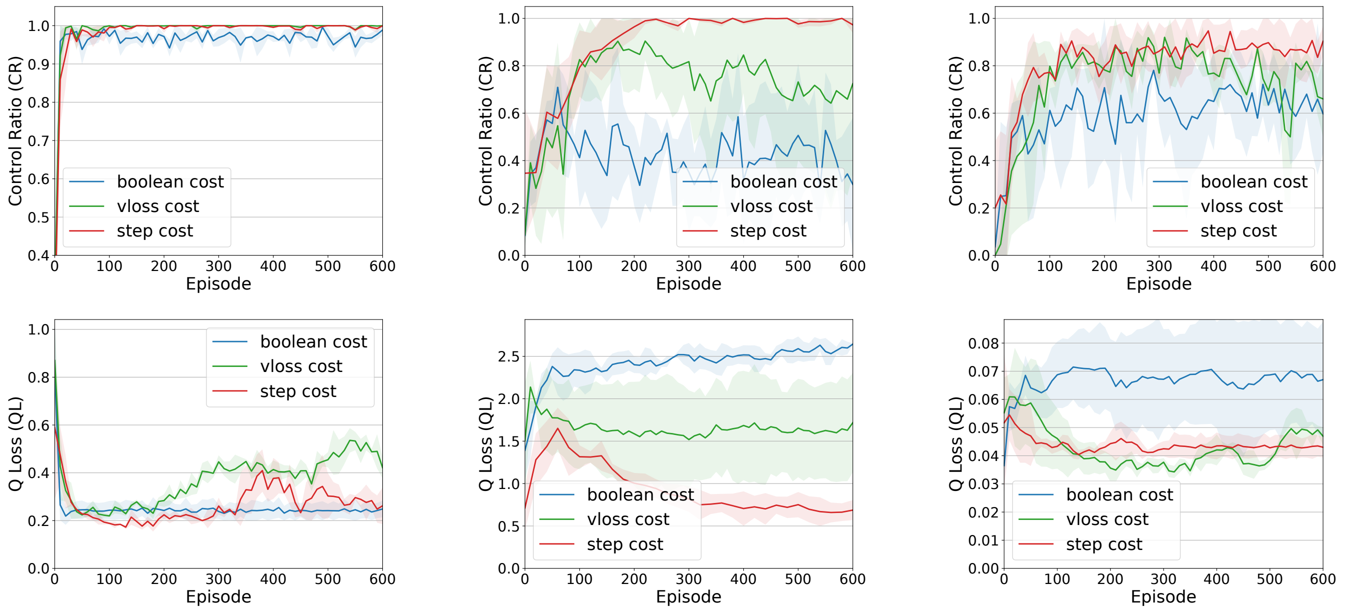
As illustrated in the figure, our proposed approach (MA-DELC) outperforms the other baseline methods in CR in all of the 3 scenarios. Specifically, MADDPG achieves a CR of approximately 0.92 in the 33-bus scenario and around 0.9 in the 141-bus scenario. COMA's and MATD3's performance is comparable to MADDPG regarding CR in the 33-bus scenario but deteriorates in the 141-bus scenario. Remarkably, our proposed MA-DELC approach attains a stable CR close to 1.0 in these two scenarios. The QL of MA-DELC is slightly higher than baselines in the 33-bus scenario, but we consider it acceptable for maintaining an outstandingly stable CR. This achievement underscores the efficacy of the Lagrangian constraint method in consistently satisfying the constraints while optimizing the objective function.

In the more complex 322-bus scenario, all baselines exhibits unstable CR and low QL of 0.03-0.05. Among the baselines, MADDPG maintains relatively higher CR (around 0.75) and lowest QL (around 0.03). MA-DELC outperforms other baselines in terms of CR, and achieves a QL close to COMA. The results of CR of MA-DELC and other baselines all exhibit high variance. We assume that in the 322-bus scenario, the intricate topological structure of the power grid demands more sophisticated agent coordination, and achieving a CR of 1.0 may be particularly challenging or even infeasible. Additionally, according to the results in Shi *et al.* [2023], the safety layer method reaches CR of 0.95 , 0.95, 0.85 in the 3 scenarios, while our method reaches a higher CR of 1.0, 1.0 and 0.9 respectively apart from reducing power loss. Overall, our results highlight the promising performance of MA-DELC, especially in scenarios with high complexity and stringent constraints, when applying to real-world.

5.3 Ablation Studies

We conduct ablation experiments to systematically assess the impact of each component within our proposed method, as well as the effects of different cost functions.

Effect of Each Component. The results of our ablation experiments are depicted in Figure 4. MA-DELC w/o cost-critic uses the single step cost estimator in training the actor network, and MA-DELC w/o cost-estimator uses the cost-critic



(a) 33-bus scenario

(b) 141-bus scenario

(c) 322-bus scenario

Figure 5: Control Ratio (CR) and Q Loss (QL) results of the different cost functions (CR: higher is better, QL: lower is better).

to update the Lagrange multiplier α. MA-DELC w/o Q-loss uses only the cost function in training.

MA-DELC w/o cost-critic exhibits lower CR in all of the three scenarios, which shows that long-term planning is necessary for satisfying the constraints in active voltage control problem. On the other hand, MA-DELC w/o cost-estimator can achieve high CR in the middle of the training, while fails to maintain it due to optimistic overestimation of cost-critic. These results validate the necessity of the proposed double-estimation of Lagrangian constraint approach.

MA-DELC w/o Q-loss exhibits clearly higher QL, which verifies that MA-DELC can fulfill the constraints and optimize the objective concurrently. Notably, MA-DELC has higher CR than MA-DELC w/o Q-loss in the 141-bus and 322-bus scenario. We assume that it is because maintaining the constraint fulfillment and optimizing the objective do not always present the opposite gradient, and Q-loss provides agents with useful information about the environment as well.

Comparisons for Cost Functions. As shown in Figure 5, we explore alternative cost functions and provide the comparative performance in terms of CR and QL. In general, the step cost function stands out as the most effective choice, outperforming the other cost functions in both CR and QL. These results echo our initial argument that cost functions with more information facilitate exploration for the agents.

When employing boolean cost functions, MA-DELC manages to learn policies achieving a CR of 1.0 in the 33-bus scenario, but fails in the 141-bus and 322-bus scenarios. This discrepancy could be attributed to the limited informativeness of the boolean cost function, which may not provide sufficient guidance for the agents in the complex large scale scenario. On the other hand, MA-DELC with the v-loss cost function exhibits unsatisfactory CR in the 141-bus scenario, but is close to the step cost function in the 322-bus scenario. The v-loss cost function, while offering more information, seems to blur the boundary of the constraint fulfillment. In the 322-bus scenario, the v-loss cost function provides additional information that can be more helpful, as the constraints are rarely fully met, leading to better performance.

These insights into the behavior of different cost functions help us better understand their impact on the learning process and the ability of the algorithm to satisfy constraints effectively. Ultimately, the choice of cost function should be considered based on the specific characteristics and requirements of the power grid scenario.

6 Conclusion

In this paper, we focus on the active voltage control problem in power distribution networks installed with PVs. We propose MA-DELC, an innovative constrained MARL algorithm, with the double estimation framework and an adaptive Lagrange multiplier, to strike a balance between minimizing power loss and ensuring safety compliance. We conducted extensive experiments and comprehensive ablation studies within real-world-scale scenarios provided by the MAPDN dataset. Experimental results demonstrate that MA-DELC with appropriate cost functions outperforms other baseline approaches. Although motivated by the AVC problem, our method is model-free and does not need any domain-specific information about the physical system. In the future, we consider testing the algorithm in more complicated settings with real-world power systems, and applying our method to other real-world safety-constrained problems.

Acknowledgments

This work was supported in part by the Major Research Plan of the National Natural Science Foundation of China (Grant No. 92048301) and Anhui Provincial Natural Science Foundation (Grant No. 2208085MF172).

References

[Achiam *et al.*, 2017] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 22–31. PMLR, 06–11 Aug 2017.

[Ackermann *et al.*, 2019] Johannes Ackermann, Volker Gabler, Takayuki Osa, and Masashi Sugiyama. Reducing overestimation bias in multi-agent domains using double centralized critics, 2019.

[Altman and Shwartz, 1998] Eitan Altman and Adam Shwartz. Constrained markov games: Nash equilibria. *Annals of Dynamic Games*, 5, 02 1998.

[Bertsekas, 2014] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic Press, 2014.

[Canese *et al.*, 2021] Lorenzo Canese, Gian Carlo Cardarilli, Luca Di Nunzio, Rocco Fazzolari, Daniele Giardino, Marco Re, and Sergio Spanò. Multi-agent reinforcement learning: A review of challenges and applications. *Applied Sciences*, 11(11), 2021.

[Cao *et al.*, 2020a] Di Cao, Weihao Hu, Junbo Zhao, Qi Huang, Zhe Chen, and Frede Blaabjerg. A multi-agent deep reinforcement learning based voltage regulation using coordinated pv inverters. *IEEE Transactions on Power Systems*, 35(5):4120–4123, 2020.

[Cao *et al.*, 2020b] Di Cao, Junbo Zhao, Weihao Hu, Fei Ding, Qi Huang, and Zhe Chen. Distributed voltage regulation of active distribution system based on enhanced multi-agent deep reinforcement learning, 2020.

[Cao *et al.*, 2021] Di Cao, Junbo Zhao, Weihao Hu, Fei Ding, Qi Huang, Zhe Chen, and Frede Blaabjerg. Data-driven multi-agent deep reinforcement learning for distribution system decentralized voltage control with high penetration of pvs. *IEEE Transactions on Smart Grid*, 12(5):4137–4150, 2021.

[Chen *et al.*, 2022] Xin Chen, Guannan Qu, Yujie Tang, Steven Low, and Na Li. Reinforcement learning for selective key applications in power systems: Recent advances and future challenges. *IEEE Transactions on Smart Grid*, 13(4):2935–2958, 2022.

[Farivar *et al.*, 2013] Masoud Farivar, Lijun Chen, and Steven Low. Equilibrium and dynamics of local voltage control in distribution systems. In *Proceedings of the 52nd IEEE Conference on Decision and Control (CDC)*, pages 4329–4334, 2013.

[Foerster *et al.*, 2017] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32, 05 2017.

[Gan *et al.*, 2013] Lingwen Gan, Na Li, Ufuk Topcu, and Steven H. Low. Optimal power flow in tree networks. In *Proceedings of the 52nd IEEE Conference on Decision and Control (CDC)*, pages 2313–2318, 2013.

[Gao *et al.*, 2021] Yuanqi Gao, Wei Wang, and Nanpeng Yu. Consensus multi-agent reinforcement learning for volt-var control in power distribution networks. *IEEE Transactions on Smart Grid*, 12(4):3594–3604, 2021.

[Gu *et al.*, 2022] Shangding Gu, Longyu Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll. A review of safe reinforcement learning: Methods, theory and applications. *ArXiv*, abs/2205.10330, 2022.

[Ha *et al.*, 2020] Sehoon Ha, Peng Xu, Zhenyu Tan, Sergey Levine, and Jie Tan. Learning to walk in the real world with minimal human effort, 2020.

[Hu *et al.*, 2022] Daner Hu, Zhenhui Ye, Yuanqi Gao, Zuzhao Ye, Yonggang Peng, and Nanpeng Yu. Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization. *IEEE Transactions on Smart Grid*, 13(6):4873–4886, 2022.

[Jahangiri and Aliprantis, 2013] Pedram Jahangiri and Dionysios C. Aliprantis. Distributed volt/var control by pv inverters. *IEEE Transactions on Power Systems*, 28(3):3429–3439, 2013.

[Liang *et al.*, 2018] Qingkai Liang, Fanyu Que, and Eytan Modiano. Accelerated primal-dual policy optimization for safe reinforcement learning, 2018.

[Liu and Wu, 2021] Haotian Liu and Wenchuan Wu. Online multi-agent reinforcement learning for decentralized inverter-based volt-var control. *IEEE Transactions on Smart Grid*, 12(4):2980–2990, 2021.

[Liu *et al.*, 2023] Zemin Eitan Liu, Quan Zhou, Yanfei Li, Shijin Shuai, and Hongming Xu. Safe deep reinforcement learning-based constrained optimal control scheme for hev energy management. *IEEE Transactions on Transportation Electrification*, 9(3):4278–4293, 2023.

[Lowe *et al.*, 2017] Ryan Lowe, YI WU, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Ma and Wu, 2023] Jinming Ma and Feng Wu. Learning to coordinate traffic signals with adaptive network partition. *IEEE Transactions on Intelligent Transportation Systems (IEEE T-ITS)*, pages 1–12, 2023.

[Pu *et al.*, 2021] Yuan Pu, Shaochen Wang, Rui Yang, Xin Yao, and Bin Li. Decomposed soft actor-critic method for cooperative multi-agent reinforcement learning, 2021.

[Ray *et al.*, 2019] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7(1):2, 2019.

[Schardt and te Heesen, 2021] Johannes Schardt and Henrik te Heesen. Performance of roof-top pv systems in selected european countries from 2012 to 2019. *Solar Energy*, 217:235–244, 2021.

[Shi *et al.*, 2022] Yuanyuan Shi, Guannan Qu, Steven Low, Anima Anandkumar, and Adam Wierman. Stability constrained reinforcement learning for real-time voltage control. In *2022 American Control Conference (ACC)*, pages 2715–2721, 2022.

[Shi *et al.*, 2023] Yufeng Shi, Mingxiao Feng, Minrui Wang, Wengang Zhou, and Houqiang Li. Multi-agent reinforcement learning with safety layer for active voltage control. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, AAMAS '23, page 1533–1541, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems.

[Sun and Qiu, 2021] Xianzhuo Sun and Jing Qiu. Two-stage volt/var control in active distribution networks with multi-agent deep reinforcement learning method. *IEEE Transactions on Smart Grid*, 12(4):2903–2912, 2021.

[Till *et al.*, 2020] Jeremy Till, Shutang You, Yilu Liu, and Pengwei Du. Impact of high pv penetration on voltage stability. In *2020 IEEE/PES Transmission and Distribution Conference and Exposition*, pages 1–5, 2020.

[Wang and Wu, 2020] Yixiang Wang and Feng Wu. Policy adaptive multi-agent deep deterministic policy gradient. In *Proceedings of the 23rd International Conference on Principles and Practice of Multi-Agent Systems (PRIMA)*, Nagoya, Japan, Novermber 2020.

[Wang *et al.*, 2019] Wei Wang, Nanpeng Yu, Jie Shi, and Yuanqi Gao. Volt-var control in power distribution systems with deep reinforcement learning. In *2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 1–7, 2019.

[Wang *et al.*, 2020] Shengyi Wang, Jiajun Duan, Di Shi, Chunlei Xu, Haifeng Li, Ruisheng Diao, and Zhiwei Wang. A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning. *IEEE Transactions on Power Systems*, 35(6):4644–4654, 2020.

[Wang *et al.*, 2021] Jianhong Wang, Wangkun Xu, Yunjie Gu, Wenbin Song, and Tim C Green. Multi-agent reinforcement learning for active voltage control on power distribution networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 3271–3284. Curran Associates, Inc., 2021.

[Wang *et al.*, 2022] Minrui Wang, Mingxiao Feng, Wengang Zhou, and Houqiang Li. Stabilizing voltage in power distribution networks via multi-agent reinforcement learning with transformer. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, KDD '22, page 1899–1909, New York, NY, USA, 2022. Association for Computing Machinery.

[Zheng *et al.*, 2016] Weiye Zheng, Wenchuan Wu, Boming Zhang, Hongbin Sun, and Yibing Liu. A fully distributed reactive power optimization and control method for active distribution networks. *IEEE Transactions on Smart Grid*, 7(2):1021–1033, 2016.