Multi-Agent Reinforcement Learning for Active Voltage Control on Power Distribution Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper presents a problem in power networks that creates an exciting and yet challenging real-world scenario for application of multi-agent reinforcement learning (MARL). The emerging trend of decarbonisation is placing excessive stress on power distribution networks. Active voltage control is seen as a promising solution to relieve power congestion and improve voltage quality without extra hardware investment, taking advantage of the controllable apparatuses in the network, such as roof-top photovoltaics (PVs) and static var compensators (SVCs). These controllable apparatuses appear in a vast number and are distributed in a wide geographic area, making MARL a natural candidate. This paper formulates the active voltage control problem in the framework of Dec-POMDP and establishes an open-source environment. It aims to bridge the gap between the power community and the MARL community and be a drive force towards real-world applications of MARL algorithms. We also analyse the special characteristics of the active voltage control problems that cause challenges for state-of-the-art MARL approaches, and point out the potential directions to solve this challenge.

1 Introduction

Multi-agent reinforcement learning (MARL) has demonstrated impressive performances on games (e.g., Chess [1], StarCraft [2, 3], and etc.) and robotics [4]. There are now increasing interests and thrusts to apply MARL in real-world problems. Power network is a natural test field for MARL algorithms. There are many problems in power networks involving the collaborative or competitive actions of a vast number of agents, such as market bidding, voltage control, frequency control, and emergency handling [5]. One of the obstacles for applying MARL in power networks is the lack of transparency and guarantee (i.e., crucial for reliable power supply for the whole society). It is sensible to start with problems that are less sensitive to reliability but hard to be solved by conventional methods.

Active voltage control in power distribution networks is such a candidate problem. The voltage control problem has been studied for years but only comes under spot light recently due to the increasing penetration of distributed resources in distribution networks, such as roof-top photovoltaics (PVs) [6]. The excessive active power injection may cause voltage fluctuations beyond the threshold of grid standards [7]. The voltage fluctuations can be relieved by exploiting the control flexibility's of PV inverters themselves along with other controllable apparatuses, such as static var compensators (SVCs) and on load tap changers

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

(OLTCs). An elaborate scheme is needed to coordinate these multiple apparatuses at scale with limited information to regulate voltage throughout the network, which is called active voltage control [8–11]. The active voltage control problem has many interesting properties. (1) It is a combination of local and global problem, that is, the voltage of each node is affected by the powers (real and reactive) of all other nodes but the impact recedes with increasing distance between nodes. (2) It is a constrained optimisation problem where the constraint is the voltage threshold and the reward is the total power loss, but there are no explicit relationship between the constraint (voltage) and the control action (reactive power). (3) A distribution network has a radial topology which contains a rich structure that can be taken as prior knowledge for control, but the detailed node-branch parameters within the topology may have significant uncertainty. (4) voltage control has a relatively large tolerance and less severe consequences if the control fails to meet standard requirements.

There have been several attempts to apply MARL in active voltage control [5, 12–15]. Each work on a particular case showed promising performances of state-of-the-art MARL approaches with modifications adapting to the active voltage control problem. However, it is not clear if these methods scale well to a larger network, and the robustness against different scenarios such as penetration level and control target, are yet to be investigated. There is no commonly accepted benchmark to provide the basis for fair comparison of different solutions.

To facilitate further research on this topic, and to bridge the gap between the power community and MARL community, we present a comprehensive test-bench and open-source environment for MARL based active voltage control. We formally define the active voltage control problem based on the framework of Dec-POMDP that is widely acknowledged in the MARL community. We write the environment in Python and construct 3 scenarios with real public data. These 3 scenarios span from small scale (i.e., 6 agents) to large scale (i.e., 38 agents). We also notice that the lack of explicit constraint in active voltage control pose a challenge for MARL. It is common in practice to convert the constraint into an auxiliary reward function [12–15] and it turns out that the performance of MARL is so sensitive to the construction of auxiliary reward, according to our results of large-scale testing.

The contributions of this paper are summarised as follows: (1) We develop an open-source environment of active voltage control for MARL and formally define this problem as a Dec-POMDP. (2) We conduct large scale experimentation with 8 state-of-the-art MARL algorithms on different scenarios of voltage control problem. (3) We notice the importance of converting voltage constraints to auxiliary reward functions and verify the sensitivity of MARL performances to the auxiliary reward functions, with a new bowl-shape auxiliary reward proposed which stands out in our test results. (4) By analysing the experimental results, we imply the possible difficulties of applying MARL and the further research directions. Through the comparison with traditional control methods, we empirically show the rationality and hope to use MARL as a solution for active voltage control.

2 Related Work

2.1 Traditional Methods for Active Voltage Control

Voltage rising and fluctuation problem in active distribution network has been studied for 20 years [16]. The traditional voltage regulating devices such as OLTC and capacitor bank [9] require a long-term schedule, so it cannot respond to intermittent renewable energy such as solar radiation in time [17]. Later works, such as [10] applied PV inverter (i.e., the device considered in this paper) on fast and continuous active voltage control. The traditional active voltage control strategies can be roughly classified into two categories: (1) reactive power dispatch based on optimal power flow (OPF) [18, 19]; and (2) droop control based on local voltage and power measurements [20, 21]. For example, centralised OPF was considered in [22–24] which can minimize the power loss while fulfilling voltage constraints. Distributed OPF was considered in [25, 26] using distributed optimization techniques, such as alternating direction method of multipliers (ADMM). The primary limitation of OPF is the need of exact network model and considerable investment in sensors [27]. Besides, solving constrained optimisation problem can be time-consuming, so it is difficult to react to the rapid change of power profile [28]. On the contrary, droop control only depends on

its local measurements. The droop control performance relies on the manually-designed parameters so that systematic stability can be hard to guarantee [28]. Consensus reactive power control was proposed in [29] while droop-based active and reactive power control was researched in [17]. However, power curtailment reduces the consumer's profit and power storage can be lossy. To solve the problems of OPF and droop control, in this paper we investigate the possibility of applying MARL techniques on controlling voltage under the rapid change of PV penetrations. Compared with the previous works, MARL is model-free and advantageous on its no need to conduct a long-term schedule and the instant reaction to the emergent change of environments.

2.2 Multi-agent Reinforcement Learning for Active Voltage Control

In this section, we discuss the work that applied MARL to active voltage control problem in the power network community. [30] applied MADDPG, with directly distributing voltage that is irregular in real world. [13, 15] applied MADDPG, with PV inverters as control variables. [12] applied MATD3, with static var compensation (SVC) and PV inverters as control variables. [14] applied MASAC, where they considered two settings of control variables, i.e., curtailment of PV, SVC and PV inverters as well as curtailment of PV, SVC and battery storage system (BSS). In these work, they abstractly modelled the problem as Markov Game with each agent managing a region of energy resources that is unrealistic in the real world. In our work, we model it as a Dec-POMDP [31] that corresponds each concept with a physical meaning in active voltage control, and each agent manages an energy resource. Moreover, the voltage losses that they leveraged exist some drawbacks, so in this paper we propose a voltage loss called Bowl loss that can reduce their weaknesses. In this paper, we concentrate on applying PV inverters to active voltage control and investigate its extreme capability for MARL with large scale experimentation.

3 Background

3.1 Power Distribution Network

The modern electric power network is illustrated by Figure 1 stage a to c. The electricity is generated from power plant and transmitted through transmission lines. Muti-stage transformers are needed to reduce the voltage levels when the electricity is delivered to the distribution network and finally consumed by residential and industrial clients. A typical PV unit consists of PV panels and voltage-source inverter which can be installed either on roof-top or in the solar farm. By convention, there exits several managing entity monitoring and operating the PV resources through the local communication channels (e.g., the smart meters). With the emerging PV penetrations, the distribution network gradually becomes an active participant in modern power networks that can deliver power to its users and the main grid (see the bidirectional power flows in Figure 1 stage d).

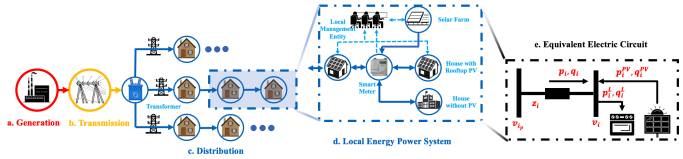


Figure 1: Illustration on distribution network (block a-b-c) under PV penetration. The solid and dotted lines represent the power and information flows respectively. Block d is the detailed version of c and block e is the electric abstract of block d (more in Appendix A.2)

.

System Model and Voltage Deviation. In this paper, we consider medium (10-24kV) and low (0.23-1kV) voltage distribution networks where the PVs are highly penetrated but cannot be efficiently controlled by the main grid. We model the distribution network in Figure 1 as a tree graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{0, 1, \ldots, N\}$ and $\mathcal{E} = \{1, 2, \ldots, N\}$ represent the set of vertices (buses) and edges (branches) respectively [22]. Bus 0 is considered as the connection to the main grid, balancing the active and reactive power in the distribution

network. For every bus $i \in \mathcal{V}$, let v_i and θ_i be the magnitude and phase angle of the complex voltage and $s_j = p_i + jq_i$ be the complex power injection. Then the active and reactive power injections can be defined as follows [32]:

$$p_i = v_i^2 \sum_{j \in \mathcal{V}_i} g_{ij} - v_i \sum_{j \in \mathcal{V}_i} v_j \left(g_{ij} \cos\theta_{ij} + b_{ij} \sin\theta_{ij} \right), \quad \forall i \in \mathcal{V} \setminus 0 \tag{1a}$$

$$q_i = -v_i^2 \sum_{j \in \mathcal{V}_i} b_{ij} + v_i \sum_{j \in \mathcal{V}_i} v_j \left(g_{ij} \sin\theta_{ij} + b_{ij} \cos\theta_{ij} \right), \quad \forall i \in \mathcal{V} \setminus 0 \tag{1b}$$

where $\mathcal{V}_i \coloneqq \{j \mid (i,j) \in \mathcal{E}\}$ is the index set of buses connected to bus i. g_{ij} and b_{ij} are the conductance and susceptance on branch (i,j). $\theta_{ij} = \theta_i - \theta_j$ is the phase difference between bus i and j. Eq.(1), which represents the environmental dynamics, is essential for solving power flow (details in Appendix A.1) and active voltage control problem. For safe and optimal operation, 5% voltage deviation is usually allowed, i.e., $0.95 \leq v_i \leq 1.05, \forall i \in \mathcal{V} \setminus 0$ and $v_0 = 1.0$ per unit. When the load is heavy during the nighttime, the end-user voltage can be smaller than 0.95 p.u.. In contrast, to export its power, large penetration of p_i^{PV} causes reverse current flow, increasing v_i out of the nominal range (Figure 1-d) [16, 33].

Optimal Power Flow. In this paper, OPF is regarded as an optimization problem minimizing on the total line power loss subject to the power balance constraints (1), PV reactive power limits, and bus voltage limits [34]. As the centralized OPF has full access to the system topology, measurements, and PV resources, it provides the optimal active voltage control performance and can be used as a benchmark method (details in Appendix A.3).

Droop Control. To regulate local voltage deviation, the standard droop control defines a piece-wise linear relationship between PV reactive power generation and voltage deviation at a bus equipped with inverter-based PVs [28, 35] (details in Appendix A.4).

3.2 Dec-POMDP

Multi-agent reinforcement learning (MARL) is an extension from the reinforcement learning problem, with multiple agents in the same environment. The problem of MARL for cooperation among agents is conventionally formulated as a Dec-POMDP [31]. It is usually formulated as a tuple such that $\langle \mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, r, \Omega, \rho, \gamma \rangle$. $\mathcal{I}$ is an agent set; $\mathcal{S}$ is a state set; $\mathcal{A} = \times_{i \in \mathcal{I}} \mathcal{A}_i$ is the joint action set, where $\mathcal{A}_i$ is each agent's action set; $\mathcal{O} = \times_{i \in \mathcal{I}} \mathcal{O}_i$ is the joint observation set, where $\mathcal{O}_i$ is each agent's observation set; $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is a transition probability function that describes the dynamics of an environment; $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a global reward function that describes the award to the whole agents given their decisions; $\Omega : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to [0,1]$ describes the perturbation of the observers (or sensors) for agents' joint observations over the states after decisions; $\rho : \mathcal{S} \to [0,1]$ is a probability function of initial states; and $\gamma \in (0,1)$ is a discount factor. The objective of Dec-POMDP is finding an optimal joint policy $\pi = \times_{i \in \mathcal{I}} \pi_i$ that solves $\max_\pi \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t]$.

4 Distributed Active Voltage Control Problem

4.1 Problem Formulation

For the ease of commercial operations, the entire power network is usually divided into several control regions. Each region includes several nodes (a.k.a. buses), with some number of PVs located. Each PV is alongside with an inverter which can generate reactive power, so that the voltage magnitudes of the whole nodes are possible to be controlled around a stationary value, denoted as v_{ref}. Since each region may contain several commercial electricity operators and it is difficult to convince all operators to share the operation authority in a region, we define each PV inverter as an agent that is operated by a company (see Figure 3 for illustrations). On the other hand, an operator is assumed to know the full information within the region it operates (where there are PV inverters that it can control). This is reasonable in the realistic market, since agents in the same region can gain the benefits with transparent information. Since each agent can only observe partial information of the whole power network, we can formulate it as a Dec-POMDP [31]. The problem can be mathematically described as a 10-tuple such that $\langle \mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{O}, \mathcal{T}, r, \Omega, \rho, \gamma \rangle$, where ρ is the

probability distribution for drawing the initial state and γ is the discount factor. The rest of definitions are shown as follows.

Agent Set. There is a set of agents controlling a set of PV inverters denoted as $\mathcal{I}$. Each agent is located at some node in $\mathcal{G}$ (i.e., a graph representing the power network defined as before). We define a function $g : \mathcal{I} \to \mathcal{V}$ to indicate the node where an agent is located.

Region Set. The whole power network is separated into M regions, whose union is denoted as $\mathcal{R} = \{\mathcal{R}_k \subset \mathcal{V} \mid k < M, k \in \mathbb{N}\} \subseteq \mathcal{V}$ and $\mathcal{R}_{k_1} \cap \mathcal{R}_{k_2} = \emptyset$ if $k_1 \neq k_2$. We define a function $f : \mathcal{V} \to \mathcal{R}$ that maps a node to the region where it is involved.

State and Observation Set. We denote $\mathbf{p}$ as the vector of active power and $\mathbf{q}$ as the vector of reactive power. The state set is defined as $\mathcal{S} = \mathcal{L} \times \mathcal{P} \times \mathcal{Q} \times \mathbb{V}$, where $\mathcal{L} = \{(\mathbf{p}, \mathbf{q}) : \mathbf{p}, \mathbf{q} \in (0, \infty)^{|\mathcal{V}|}\}$ is a set of (active and reactive) power of loads; $\mathcal{P} = \{\mathbf{p} : \mathbf{p} \in (0, \infty)^{|\mathcal{I}|}\}$ is a set of active power generated from PVs; $\mathcal{Q} = \{\mathbf{q} : \mathbf{q} \in (0, \infty)^{|\mathcal{I}|}\}$ is a set of reactive power generated from PV inverters at the last step; $\mathbb{V} = \{(\mathbf{v}^m, \mathbf{v}^d) : \mathbf{v}^m \in (0, \infty)^{|\mathcal{V}|}, \mathbf{v}^d \in [-\pi, \pi]^{|\mathcal{V}|}\}$ is a set of voltage wherein $\mathbf{v}^m$ is a vector of voltage magnitudes and $\mathbf{v}^d$ is a vector of voltage phases measured in radius. We define a function $h : \mathbb{P}(\mathcal{V}) \to \mathbb{P}(\mathcal{S})$ that maps a subset of $\mathcal{V}$ to its correlated measures, where $\mathbb{P}(\cdot)$ denotes the power set. The observation set is defined as $\mathcal{O} = \times_{i \in \mathcal{I}} \mathcal{O}_i$, where $\mathcal{O}_i = (h \circ f \circ g)(i)$ is a subset of $\mathcal{S}$ including the measures within the region where agent i is located.

Action Set. Each agent $i \in \mathcal{I}$ is equipped with a continuous action set $\mathcal{A}_i = \{a_i \ : \ -c \leq a_i \leq c, c > 0\}$. The continuous action represents the ratio of maximum reactive power it generates, i.e., the reactive power generated from the kth PV inverter is $\mathbf{q}[k] = a_k \sqrt{(\mathbf{s}^{\max}[k])^2 - (\mathbf{p}[k])^2}$, where $\mathbf{s}^{\max}[k]$ is the maximum apparent power of the kth node that is dependent on the physical capacity of the PV inverter. If $a_k > 0$, it means providing reactive power to the power network. If $a_k < 0$, it means absorbing reactive power from the power network. The value of c is usually selected as per the loading capacity of a power network, which is for the safety of operations. The joint action set is denoted as $\mathcal{A} = \times_{i \in \mathcal{I}} \mathcal{A}_i$.

State Transition Probability Function. Since the state includes the last action and the change of loads is random (that theoretically can be modelled as any probabilistic distribution), we can naturally define the state transition probability function as $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ that follows Markov decision process. Specifically, $\mathcal{T}(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t) = \delta(\mathbf{s}_t, \mathbf{a}_t) + Pr(\mathbf{s}_{t+1} | \mathbf{s}_t)$, where $\mathbf{a}_t \in \mathcal{A}$ and $\mathbf{s}_t, \mathbf{s}_{t+1} \in \mathcal{S}$. $\delta(\mathbf{s}_t, \mathbf{a}) \mapsto \mathbf{s}_{t+1}$ denotes the solution of the power flow, whereas $Pr(\mathbf{s}_{t+1} | \mathbf{s}_t)$ describes the change of states from loads (i.e., highly correlated to the user behaviours) with other states fixed.

Observation Probability Function. We now define the observation probability function. In the context of power network, it describes the measurement errors that may occur from sensors. Mathematically, we can define it as $\Omega : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to [0, 1]$. Specifically, $\Omega(\mathbf{o}_{t+1} | \mathbf{s}_{t+1}, \mathbf{a}_t) = \mathbf{s}_{t+1} + \mathcal{N}(\mathbf{0}, \Sigma)$, where $\mathcal{N}(\mathbf{0}, \Sigma)$ is an isotropic multi-variable Gaussian distribution and Σ is dependent on the physical properties of sensors, e.g., smart meters.

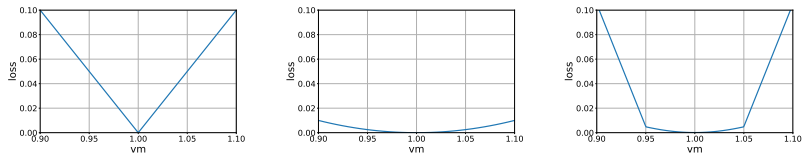
Reward Function. The reward function is defined as follows:

$$r = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} l_v(\mathbf{v}^m[i]) - \alpha \cdot l_q(\mathbf{q}), \tag{2}$$

where $l_v(\cdot)$ is a voltage loss and $l_q(\mathbf{q}) = \frac{1}{|\mathcal{I}|} ||\mathbf{q}||_1$ is the reactive power generation loss (i.e., a type of power loss approximation easy to compute). The objective is to control the voltage within a safety range around v_{ref}, while the reactive power generation is as less as possible, i.e., $l_q(\mathbf{q}) < \epsilon$ and $\epsilon > 0$. Similar to the mathematical tricks used in β-VAE [36], by KKT conditions we can transform a constrained reward to a unconstrained reward with a Lagrangian multiplier $\alpha \in (0, 1)$ shown in Eq.2 (regarding reward as an objective function with maximization). Since $l_v(\cdot)$ is not easy to define in practice (i.e., it affects $l_q(\mathbf{q})$), we aim to study for a good choice in this paper.

Objective Function. The objective function of this problem is $\max_\pi \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t]$, where $\pi = \times_{i \in \mathcal{I}} \pi_i$; $\pi_i : \bar{\mathcal{O}}_i \times \mathcal{A}_i \to [0, 1]$ and $\bar{\mathcal{O}}_i = (\mathcal{O}_i^\tau)_{\tau=1}^h$ is the history of observations with history length as h. Literally, we need to find an optimal joint policy π to maximize the discounted cumulative rewards.

4.2 Voltage Loss



(a) L1 loss. (b) L2 loss. (c) Bowl loss.

Figure 2: This figure shows 3 voltage losses, where L1 loss and L2 loss are two baseline losses while Bowl loss is proposed in this paper.

We define $v_{\text{ref}} = 1$ in this paper, and the voltage needs to be controlled within the safety range from 0.95 to 1.05. Since the voltage loss is difficult to define in practice, we investigate a suitable one in this paper. L1 loss (see Figure 2a) was most frequently used in the previous work [12–14], however, this may lead to wasteful reactive power generations since $\frac{|\Delta l_v|}{\alpha|\Delta l_q|} \gg 1$ within the safety range of voltage. Although L2 loss (see Figure 2b) used in other work [15] may alleviate this problem, it may be slow to guide the policy outside the safety range. To address these problems, we propose a loss called Bowl loss. Bowl loss combines the advantages of L1 and L2 loss. It gives a steep gradient outside the safety range, while provides a slighter gradient as voltage tends to the v_{ref} that enables $\frac{|\Delta l_v|}{\alpha|\Delta l_q|} \to 0$ as $v \to v_{\text{ref}}$.

5 Experiments

5.1 Experimental Settings

Power Network Topologies. Two MV networks, IEEE 33-bus [37] and 141-bus [38] are modified as the first two topologies under test. The original topologies and parameters can be found in MATPOWER [39]. To show the flexibility on network with multi-voltage levels, we construct a 110kV-20kV-0.4kV (high-medium-low voltage) 322-bus network using benchmark topologies from SimBench [40]. Each region consists of 1-4 PVs depending on various zonal sizes. The specific network descriptions and partitions can be found in Appendix C.1. In particular, Figure 3 illustrates the 33-bus system.

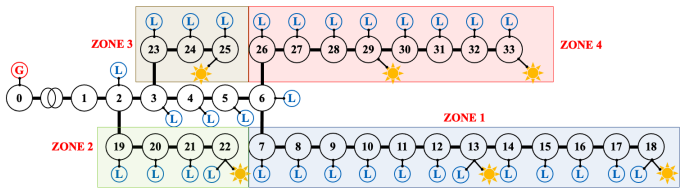
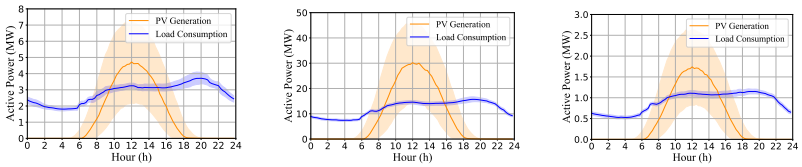


Figure 3: Illustration on 33-bus system. Each node (bus) is indexed by the large circle. Four control regions are partitioned by the smallest path from the terminal to the main branch (bus 1-6). We regulate voltage profiles on bus 2-33 whereas bus 0-1 are considered as substation with constant voltage and infinite active and reactive power capacity. G represents the external generator while the small Ls represent loads. ☀ represents the location of PVs.

Data Descriptions. The load profile of each network is modified from real-time Portuguese electricity consumption accounting for 232 consumers of 3 years [41]. To highlight the differences between residential and industrial users, we randomly perturb $\pm 5\%$ on the default power factors defined in the case files and generate real-time reactive power consumption accordingly. The solar data is collected from Elia group [42], a Belgium's power network operator. To distinguish on different solar radiation levels in different regions, the 3-year PV generations from 10 cites/regions are collected and PVs in the same control region have the same generation profiles. We define the PV penetration rate (PR) as the ratio between rated PV generation and rated load consumption. In this paper, we set $PR \in \{2.5, 3, 4\}$ for different topologies. We oversize the PV inverter by 20% of its maximum active power

generation to meet the minimum IEEE requirement [35]. The median and 25%-75% quantile shadings of loads and PV generations are illustrated in Figure 4 while the details can be found in Appendix C.2.



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

Figure 4: Active PV generation and load consumption.

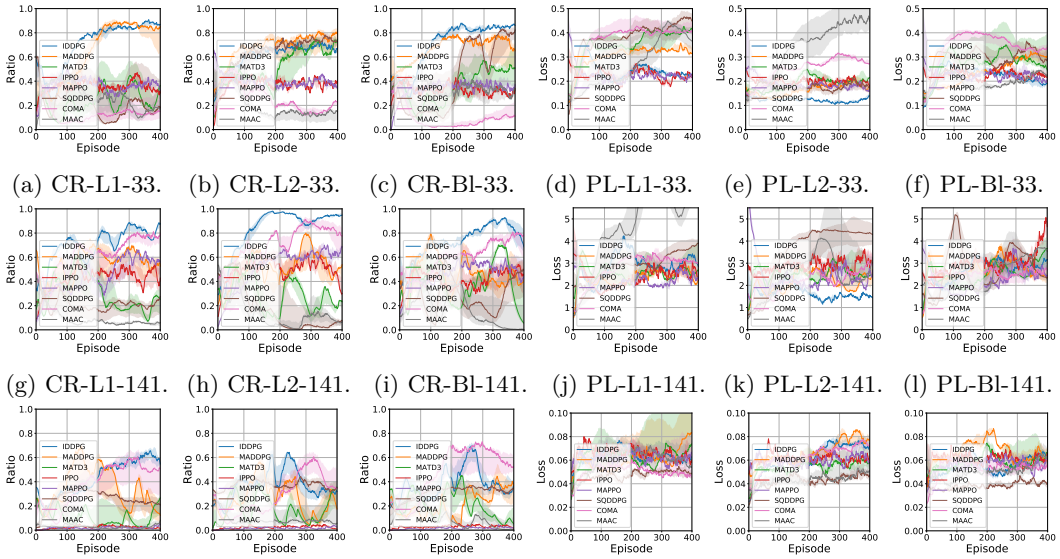
MARL Simulation Settings. We now describe the simulation settings standing by the view of MARL. In 33-bus system, there are 4 regions and 6 agents. In 141-bus system, there are 9 regions and 22 agents. In 322-bus system, there are 22 regions and 38 agents. The discount factor γ is set to 0.99. α in Eq.2 is set to 0.01. To guarantee the safety of power networks, we manually set the range of actions for each scenario, with $[-0.5, 0.5]$ for 33-bus system, $[-0.8, 0.8]$ for 141-bus system, and $[-0.3, 0.3]$ for 322-bus system. During training, we randomly sample the initial state for an episode and each episode lasts for 100 time steps (i.e., nearly a day). Every experiment is run with 3 random seeds and the test results during training are given via the median and the 25%-75% quartile shading.

Evaluation Metrics. In experiments, we apply two metrics to evaluate the performance of algorithms. **Total controllable Rate (CR):** it calculates the ratio of all buses' voltage being under control for each episode. **Average power loss (PL):** It calculates the average power loss of all buses for each episode. We aim to find algorithms and reward functions with high CR and low PL.

MARL Algorithm Settings. We evaluate the performances of state-of-the-art MARL algorithms, i.e., IDDPG [43], MADDPG [44], COMA [45], IPPO [2], MAPPO [3], MAAC [46], SQDDPG [43] and MATD3 [47] on this real-world problem with continuous actions. Since original COMA can only work for discrete actions, we conduct some modifications so that it can work for continuous actions (see Appendix B). All algorithms are implemented following their open source codes and the details posted on their papers. The neural networks of all policies and critics are respectively equipped with a 3-layer MLPs, with a 32-dim hidden layer. The activation functions for first two layers are ReLUs for all algorithms except for MAPPO that is suggested using Tahn as activation functions. The input of policies and critics is the concatenation of the current observation and the previous observation. We use RMSProp [48] as the optimizer, with the learning rate of 1e-4 for updating policy and 1e-3 for updating critic respectively.

5.2 Main Results

We first show the main results of all algorithms on all scenarios in Figure 5. It is surprising that IDDPG generally performs best on most scenarios (e.g., bus-33 with L1, bus-33 with Bowl, bus-141 with L1, bus-141 with L2, bus-322 with L1), which is consistent to its good performance reported from [43]. MADDPG performs well on bus-33, however, when the system scale becomes larger and the number of agents increases, its performance drastically falls. In contrast, the performance of COMA is not pleasant when the number of agents is small, while it performs well on the system with the large number of agents. Although MAPPO and IPPO performed well on games [2, 3], their performances on the real-world power network problems are poor. The most probable reason could be that the variations of dynamics and uncertainties in power networks are far faster and more complicated than games and their conservative policy updates cannot fast respond. MATD3 and MAAC perform poorly on any scenarios and the reasons need to be investigated. SQDDPG performs well on some scenarios (e.g., bus-33 with L2, bus-33 with Bowl, and bus-322 with L2), but not stable on most of scenarios. The main reason could be its overgeneralisation and instability caused by its permutations when constructing a critic [43].

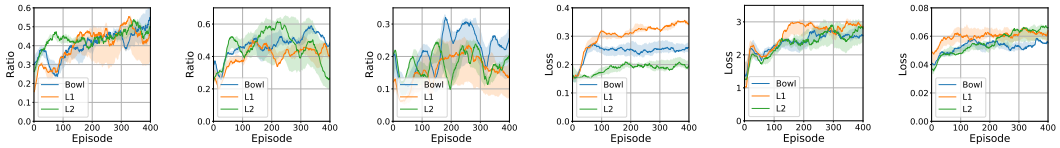


(a) CR-L1-33. (b) CR-L2-33. (c) CR-Bl-33. (d) PL-L1-33. (e) PL-L2-33. (f) PL-Bl-33.

(g) CR-L1-141. (h) CR-L2-141. (i) CR-Bl-141. (j) PL-L1-141. (k) PL-L2-141. (l) PL-Bl-141.

(m) CR-L1-322. (n) CR-L2-322. (o) CR-Bl-322. (p) PL-L1-322. (q) PL-L2-322. (r) PL-Bl-322.

Figure 5: Performance of algorithms with different voltage losses. The sub-caption indicates metric-voltage loss-scenario and Bl is the contraction of Bowl.



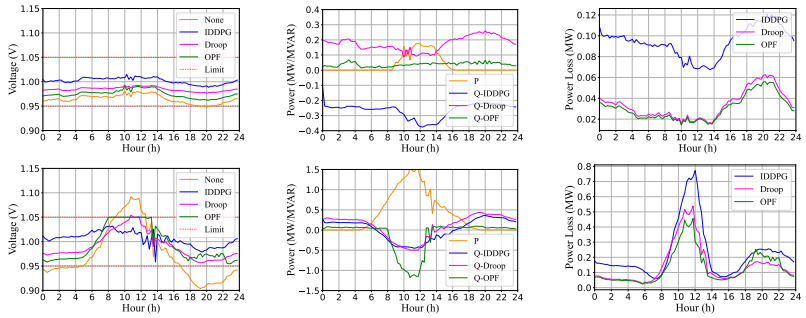
(a) CR-33. (b) CR-141. (c) CR-322. (d) PL-33. (e) PL-141. (f) PL-322.

Figure 6: Median performance of overall algorithms with different voltage losses. The sub-caption indicates metric-scenario.

To study the effects of different voltage losses, we show the median performance of overall 8 MARL algorithms in Figure 6. It can be observed that Bowl loss proposed in this paper can almost work well for all algorithms on all scenarios, except for the bus-33 where its PL is a bit larger than L2. Among these 3 voltage losses, L1 (i.e., the one most used in the past work [13, 12, 14]) is the worst choice for the active voltage control, since it always results in superfluous power loss that is in conflict with the power network optimal operations [49, 34]. L2 is a good choice for the small scale power network with proper CR and low PL, while it performs worse on 141-bus with too low CR and 322-bus with too high PL.

5.3 Relationship between MARL and Traditional Control Methods

To further study the relationship between MARL algorithms and traditional control methods, we conduct a case study on 33-bus with IDDPG and Bowl loss as the candidate for MARL. The traditional control methods we select here are OPF [22] and droop control [20]. To ease life, we only demonstrate a bus with PV (i.e., one of the most difficult buses to control) on one day in summer and winter respectively. It is apparent from Figure 7a and 7b that the policy of IDDPG is similar to droop control. Although the reactive power of IDDPG is overly generated in winter with a large bias, the trend is still similar. Compared with OPF, the reactive power generations of IDDPG generally overflow and therefore its power loss becomes larger than OPF's. The main reason could be that MARL algorithms cannot get the information of system model and its control frequency is far lower than OPF's. As a result, MARL algorithms have to make decision more safely in case of the rapid variations of environments, e.g., cloud cover.



(a) Voltage. (b) Power. (c) Power Loss.

Figure 7: Compare IDDPG with traditional control methods on bus 18 of 33-bus system. The first row: a winter day. The second row: a summer day. None and limit in the first column represent the voltage without control and the safety range respectively. P and Q in the second column indicate the PV active and reactive power generated via different methods respectively.

5.4 Discussion

We now discuss the results that we obtained from experimentation. First, for the active voltage control problem the performances of current MARL algorithms are so sensitive to the voltage loss. Through our experimental findings, we wrap up two possible suggestions for the further MARL research on active voltage control: (1) the MARL algorithms should be studied along with a voltage loss rather than solely developed; (2) the study for the reward function design should be recognized, since this may influence the best performances of MARL algorithms to active voltage control even with tiny changes.

Second, MARL algorithms actually mimic the strategy of a traditional control method called droop control that can well solve active voltage control with limitations such that (1) the restriction on the usage for the buses with PVs, (2) the need of frequent measurements (or observations) for the promising control performance, and (3) the difficulty of manually tuning parameters. This can be looked as an evidence for the rationality and possibility to address active voltage control by MARL, which is the first time to be revealed to our best knowledge. On the other hand, MARL is able to reduce the limitations of droop control and this is the reason why we need to conduct the further research.

Third, our experiments show that the difficulties of active voltage control for MARL is not only the high control performance but also the low power loss. The latter was usually omitted in the previous work [12–15], however, it is extremely important to the power network since the low power loss implies the economic operations. The main difficulty of balancing these two indices is the coupling relationship between voltage and power [10]. For this reason, to investigate reducing power loss while keeping high control performance for MARL on the active voltage control is significant.

6 Conclusion

This paper studies applying multi-agent reinforcement learning (MARL) to active voltage control in the power distribution networks. We formally define this problem as a Dec-POMDP and propose a voltage loss called Bowl loss for the reward function. Observing from large scale experimentation, we conclude the difficulties and possible research directions of MARL for active voltage control. Moreover, we compare the performance of MARL algorithms with traditional control methods and empirically show the rationality and possibility of regarding MARL as a substitute for the traditional control methods. To facilitate the further research, we develop an open-source environment and provide 3 scenarios with real-world data. In the future, we plan to investigate the MARL algorithms with explanability for this problem. For the real-world problems such as active voltage control, the explanability is as important as its performance, mainly due to the requirement of safety and trust during control process.

References

[1] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[2] Christian Schroeder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

[3] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of mappo in cooperative, multi-agent games. *arXiv preprint arXiv:2103.01955*, 2021.

[4] Christian Schroeder de Witt, Bei Peng, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. Deep multi-agent reinforcement learning for decentralized continuous cooperative control. *arXiv preprint arXiv:2003.06709*, 2020.

[5] Di Cao, Weihao Hu, Junbo Zhao, Guozhou Zhang, Bin Zhang, Zhou Liu, Zhe Chen, and Frede Blaabjerg. Reinforcement learning and its applications in modern power and energy systems: A review. *Journal of Modern Power Systems and Clean Energy*, 8(6): 1029–1042, 2020. doi: 10.35833/MPCE.2020.000552.

[6] Efstratios Batzelis, Zakir Hussain Rather, John Barton, Bonu Ramesh Naidu, Billy Wu, Firdous Ul Nazir, ONYEMA SUNDAY NDUKA, Wei He, Jerome Nsengiyaremye, and Bandopant Pawar. Solar integration in the uk and india: technical barriers and future directions. *Report of the UK-India Joint Virtual Clean Energy Centre (JVCEC)*, 2021.

[7] Reinaldo Tonkoski, Dave Turcotte, and Tarek H. M. EL-Fouly. Impact of high pv penetration on voltage profiles in residential neighborhoods. *IEEE Transactions on Sustainable Energy*, 3(3):518–527, 2012. doi: 10.1109/TSTE.2012.2191425.

[8] Pedro MS Carvalho, Pedro F Correia, and Luís AFM Ferreira. Distributed reactive power generation control for voltage rise mitigation in distribution networks. *IEEE transactions on Power Systems*, 23(2):766–772, 2008.

[9] Tomonobu Senjyu, Yoshitaka Miyazato, Atsushi Yona, Naomitsu Urasaki, and Toshihisa Funabashi. Optimal distribution voltage control and coordination with distributed generation. *IEEE Transactions on Power Delivery*, 23(2):1236–1242, 2008. doi: 10.1109/TPWRD.2007.908816.

[10] Konstantin Turitsyn, Petr Sulc, Scott Backhaus, and Michael Chertkov. Options for control of reactive power by distributed photovoltaic generators. *Proceedings of the IEEE*, 99(6):1063–1073, 2011. doi: 10.1109/JPROC.2011.2116750.

[11] Johanna Barr and Ritwik Majumder. Integration of distributed generation in the volt/var management system for active distribution networks. *IEEE Transactions on Smart Grid*, 6(2):576–586, 2015. doi: 10.1109/TSG.2014.2363051.

[12] 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. *arXiv preprint arXiv:2006.00546*, 2020.

[13] 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.

[14] 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*, 2021.

[15] Haotian Liu and Wenchuan Wu. Online multi-agent reinforcement learning for decentralized inverter-based volt-var control. *IEEE Transactions on Smart Grid*, 2021.

[16] C.L. Masters. Voltage rise: the big issue when connecting embedded generation to long 11 kv overhead lines. *Power Engineering Journal*, 16(1):5–12, 2002. doi: 10.1049/pe:20020101.

[17] Giuseppe Fusco and Mario Russo. A decentralized approach for voltage control by multiple distributed energy resources. *IEEE Transactions on Smart Grid*, pages 1–1, 2021. doi: 10.1109/TSG.2021.3057546.

[18] Yashodhan P. Agalgaonkar, Bikash C. Pal, and Rabih A. Jabr. Distribution voltage control considering the impact of pv generation on tap changers and autonomous regulators. *IEEE Transactions on Power Systems*, 29(1):182–192, 2014. doi: 10.1109/TPWRS.2013.2279721.

[19] Panagis N. Vovos, Aristides E. Kiprakis, A. Robin Wallace, and Gareth P. Harrison. Centralized and distributed voltage control: Impact on distributed generation penetration. *IEEE Transactions on Power Systems*, 22(1):476–483, 2007. doi: 10.1109/TPWRS.2006.888982.

[20] Pedram Jahangiri and Dionysios C. Aliprantis. Distributed volt/var control by pv inverters. *IEEE Transactions on Power Systems*, 28(3):3429–3439, 2013. doi: 10.1109/TPWRS.2013.2256375.

[21] Johannes Schiffer, Thomas Seel, Jörg Raisch, and Tevfik Sezi. Voltage stability and reactive power sharing in inverter-based microgrids with consensus-based distributed voltage control. *IEEE Transactions on Control Systems Technology*, 24(1):96–109, 2016. doi: 10.1109/TCST.2015.2420622.

[22] Lingwen Gan, Na Li, Ufuk Topcu, and Steven H Low. Optimal power flow in tree networks. In *52nd IEEE Conference on Decision and Control*, pages 2313–2318. IEEE, 2013.

[23] Yan Xu, Zhao Yang Dong, Rui Zhang, and David J Hill. Multi-timescale coordinated voltage/var control of high renewable-penetrated distribution systems. *IEEE Transactions on Power Systems*, 32(6):4398–4408, 2017.

[24] Emiliano Dall'Anese, Sairaj V. Dhople, and Georgios B. Giannakis. Optimal dispatch of photovoltaic inverters in residential distribution systems. *IEEE Transactions on Sustainable Energy*, 5(2):487–497, 2014. doi: 10.1109/TSTE.2013.2292828.

[25] 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. doi: 10.1109/TSG.2015.2396493.

[26] Zhiyuan Tang, David J Hill, and Tao Liu. Distributed coordinated reactive power control for voltage regulation in distribution networks. *IEEE Transactions on Smart Grid*, 12(1):312–323, 2020.

[27] Hongbin Sun, Qinglai Guo, Junjian Qi, Venkataramana Ajjarapu, Richard Bravo, Joe Chow, Zhengshuo Li, Rohit Moghe, Ehsan Nasr-Azadani, Ujjwol Tamrakar, Glauco N. Taranto, Reinaldo Tonkoski, Gustavo Valverde, Qiuwei Wu, and Guangya Yang. Review of challenges and research opportunities for voltage control in smart grids. *IEEE Transactions on Power Systems*, 34(4):2790–2801, 2019. doi: 10.1109/TPWRS.2019.2897948.

[28] Ankit Singhal, Venkataramana Ajjarapu, Jason Fuller, and Jacob Hansen. Real-time local volt/var control under external disturbances with high pv penetration. *IEEE Transactions on Smart Grid*, 10(4):3849–3859, 2019. doi: 10.1109/TSG.2018.2840965.

[29] Mehdi Zeraati, Mohamad Esmail Hamedani Golshan, and Josep M Guerrero. Voltage quality improvement in low voltage distribution networks using reactive power capability of single-phase pv inverters. *IEEE transactions on smart grid*, 10(5):5057–5065, 2018.

[30] 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.

[31] Frans A Oliehoek and Christopher Amato. *A concise introduction to decentralized POMDPs*. Springer, 2016.

[32] Hadi Saadat et al. *Power system analysis*, volume 2. McGraw-hill, 1999.

[33] Guangya Yang, Francesco Marra, Miguel Juamperez, Søren Bækhøj Kjær, Seyedmostafa Hashemi, Jacob Østergaard, Hans Henrik Ipsen, and Kenn H. B. Frederiksen. Voltage rise mitigation for solar pv integration at lv grids studies from pvnet. dk. *Journal of Modern Power Systems and Clean Energy*, 3(3):411–421, 2015. doi: 10.1007/s40565-015-0132-0.

[34] Antonio Gómez-Expósito, Antonio J Conejo, and Claudio Cañizares. *Electric energy systems: analysis and operation*. CRC press, 2018.

[35] Ieee standard for interconnection and interoperability of distributed energy resources with associated electric power systems interfaces. *IEEE Std 1547-2018 (Revision of IEEE Std 1547-2003)*, pages 1–138, 2018. doi: 10.1109/IEEESTD.2018.8332112.

[36] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[37] M.E. Baran and F.F. Wu. Network reconfiguration in distribution systems for loss reduction and load balancing. *IEEE Transactions on Power Delivery*, 4(2):1401–1407, 1989. doi: 10.1109/61.25627.

[38] H.M. Khodr, F.G. Olsina, P.M. De Oliveira-De Jesus, and J.M. Yusta. Maximum savings approach for location and sizing of capacitors in distribution systems. *Electric Power Systems Research*, 78(7):1192–1203, 2008. ISSN 0378-7796. doi: https://doi.org/10.1016/j.epsr.2007.10.002. URL `https://www.sciencedirect.com/science/article/pii/S0378779607002143`.

[39] Ray Daniel Zimmerman, Carlos Edmundo Murillo-Sánchez, and Robert John Thomas. Matpower: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on power systems*, 26(1):12–19, 2010.

[40] Steffen Meinecke, Džanan Sarajlić, Simon Ruben Drauz, Annika Klettke, Lars-Peter Lauven, Christian Rehtanz, Albert Moser, and Martin Braun. Simbench—a benchmark dataset of electric power systems to compare innovative solutions based on power flow analysis. *Energies*, 13(12), 2020. ISSN 1996-1073. URL `https://www.mdpi.com/1996-1073/13/12/3290`.

[41] Artur Trindade. Electricity load diagrams 2011-2014 data set, 2015. URL `https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014`.

[42] Elia. Solar power generation, 2021. URL `https://www.elia.be/en/grid-data/power-generation/solar-pv-power-generation-data`.

[43] Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. Shapley q-value: A local reward approach to solve global reward games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7285–7292, Apr 2020.

[44] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6379–6390, 2017.

[45] Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[46] Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 2961–2970. PMLR, 2019.

[47] Johannes Ackermann, Volker Gabler, Takayuki Osa, and Masashi Sugiyama. Reducing overestimation bias in multi-agent domains using double centralized critics. *arXiv preprint arXiv:1910.01465*, 2019.

[48] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[49] Vassilis Kekatos, Gang Wang, Antonio J. Conejo, and Georgios B. Giannakis. Stochastic reactive power management in microgrids with renewables. *IEEE Transactions on Power Systems*, 30(6):3386–3395, 2015. doi: 10.1109/TPWRS.2014.2369452.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [N/A] This is not suitable for this work, since this work proposes a new real-world problem rather than an algorithm.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] The real-world problem discussed in this paper has no negative societal impacts up to now.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 4.1.
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] see Section 5.1 and Appendix D.1.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] see Section 5.1 and Appendix D.2.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] see Section 5.2.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] see Appendix D.2.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] see Section 5.1.

(b) Did you mention the license of the assets? [Yes] see Section 5.1.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] see the related references in Section 5.1.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] All data used in this work is public available.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] All data used in this work is public available.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]