Stability Constrained Reinforcement Learning for Real-Time Voltage Control

Yuanyuan Shi[1,*], Guannan Qu[2,*], Steven Low[3], Anima Anandkumar[3] and Adam Wierman[3]

***Abstract*—Deep reinforcement learning (RL) has been recognized as a promising tool to address the challenges in real-time control of power systems. However, its deployment in real-world power systems has been hindered by a lack of formal stability and safety guarantees. In this paper, we propose a stability constrained reinforcement learning method for real-time voltage control in distribution grids and we prove that the proposed approach provides a formal voltage stability guarantee. The key idea underlying our approach is an explicitly constructed Lyapunov function that certifies stability. We demonstrate the effectiveness of the approach in case studies, where the proposed method can reduce the transient control cost by more than 30% and shorten the response time by a third compared to a widely used linear policy, while always achieving voltage stability. In contrast, standard RL methods often fail to achieve voltage stability.**

***Index Terms*—reinforcement learning, Lyapunov stability, voltage control**

I. Introduction

To achieve high penetration of distributed renewable generations at the distribution grid level, maintaining the voltage levels within the safe limit has been increasingly challenging. Recently, significant efforts have been put into the design of real-time feedback controllers for voltage control purposes, see e.g. [1]–[9]. Despite the progress, most of the existing work has only been able to optimize the steady state cost, i.e. the cost of the operating point after the voltage converges into the safe limit. In the meanwhile, transient performance is of equal importance. For example, once voltage violation happens, an important goal is to bring the voltage profile back to the safe limit as soon as possible, or in other words, to minimize the voltage recovery time. However, optimizing or even analyzing the transient cost like the voltage recovery time has long been challenging as this is a nonlinear control problem. The challenge has further been complicated by the fact that many existing works [3], [5]–[9] require knowledge of the topology and the parameters of the electric grid. Yet, such knowledge is not always available due to frequent system reconfigurations (resulting in topology change [10]) and time-varying system parameters resulting from facility aging, temperature and humidity change, etc.

Reinforcement Learning (RL), having achieved impressive success in the past decade in game play [11], robotics [12], is a promising tool to address the above challenges. RL methods do not need knowledge of explicit models and can learn from interactions with the underlying system. Further, due to the expressive power of neural networks as controllers/policies, RL is effective in learning nonlinear controllers with good transient performance. As a result, there has been tremendous interest in using RL for voltage control [13]–[25], see [26] for a recent review.

Despite this interest, it has been generally agreed that the key challenge in applying RL to power systems is the *stability* issue [26]. Specifically, power systems are critical infrastructure systems that place a high emphasis on stability, i.e. the ability to maintain at safe operating points under disturbances. Stability is important because instability can lead to unsafe operating conditions that violate regulatory requirements [27] or even lead to catastrophic consequences, e.g., blackouts [28]. Despite the importance of stability, off-the-shelf RL algorithms lack provable stability guarantees. In particular, popular RL methods for continuous control, such as deep deterministic policy gradient (DDPG) [29], are gradient-based methods that focus on minimizing cost and do not explicitly consider stability. Even if the learned policy may appear "stable" on the training data set, it is not guaranteed to be stable as stability is a worst case concept requiring provably checking under the worst case scenario, which off-the-shell RL methods do not consider. The lack of provable stability guarantees is one of the biggest hurdles in applying RL to power systems since as mentioned earlier, instability can be catastrophic.

Motivated by the challenge above, the question we address in this paper is: *Can we apply RL to voltage control with provable stability guarantee?*

Contributions. We answer the question affirmatively by designing a stability constrained RL framework that learns a control policy to optimize transient cost for voltage control with provable stability guarantees. The key idea underlying our approach is that we show strict monotonicity of the policy is sufficient to formally guarantee stability (Theorem 1). The technique underlying Theorem 1 is that we derive an explicitly constructed Lyapunov function, which we use to certify stability for all monotone policies. With this stability result, we propose a Stable-DDPG approach which integrates the monotone constraint with DDPG through monotone policy network design (Corollary 1). The proposed method enables us to leverage the power of RL to improve the transient performance of voltage control without knowing the underlying model parameters, and in the meanwhile provably guarantee stability during and after the training. To the best of our knowledge, this is the first RL approach that learns nonlinear

*Authors contributed equally.

[1] Yuanyuan Shi is with the Department of Electrical and Computer Engineering, University of California San Diego, `yyshi@eng.ucsd.edu`

[2] Guannan Qu is with the Department of Electrical Engineering, Carnegie Mellon University `gqu@andrew.cmu.edu`

[3] Steven Low, Anima Anandkumar and Adam Wierman are with the Computing and Mathematical Sciences, California Institute of Technology.

policies with stability guarantees for voltage control.

We also perform numerical case studies to demonstrate the effectiveness and stability of the proposed method with both simulated disturbances and real-world data. Our method guarantees voltage stability under all operating conditions, which is not true for the standard DDPG method. In addition, our method can reduce the transient control cost by more than 30% and shorten the voltage recovery time by a third compared to a widely used linear policy in the literature [1], [3].

Related Work. This paper connects to a broad set of literature in RL, control, and power systems.

Lyapunov-based Policy Learning. The Lyapunov theory is a systematic framework to analyze the stability of a control system. The core idea is to identify a positive definite function (i.e., a Lyapunov function) of the system's state, with negative derivatives along system trajectories [30]. Using Lyapunov functions in RL was first introduced by [31], but the work did not discuss how to find a candidate Lyapunov function in general except for a case-by-case construction. A set of recent works including [32]–[36] have attempted to address this challenge by jointly learning the policy and the Lyapunov function, where [34] uses linear programming and [32], [33], [35], [36] parameterizes the Lyapunov function as neural networks. In the context of these works, our contribution can be viewed as explicitly constructing a Lyapunov function for the voltage control problem and using it to guide policy learning.

Reinforcement Learning for Power Systems. Our work contributes to a growing line of papers that use RL for voltage control [13]–[25], see [26] for a recent review. As pointed out in [26], one of the key issues in RL for power system control is the lack of provable stability guarantees, and our work makes a step toward addressing this issue by providing a formal stability guarantee on the learned policy. In particular, our experiments compare against [21], which uses standard multi-agent DDPG for voltage control. Closest in spirit to our paper is [37], which proposes a stable RL approach for frequency control via a Lyapunov approach. However, their approach only applies to the frequency control application, while our method works for voltage control which requires a different Lyapunov function design. Interestingly, both our work and prior work [37] arrive at a similar stability condition, that is strict policy monotonicity guarantees system stability.

II. Preliminaries

In this section, we first introduce the distribution system power flow models and the voltage control problem formulation. Then, we review some background for policy optimization in reinforcement learning.

A. Branch Flow Model for Distribution Networks

We consider the distribution network as a tree-structured graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, which consists of a set of nodes $\mathcal{N}_0 = \{0, 1, \dots, n\}$ and edges $\mathcal{E}$, where node 0 is the substation. We use $\mathcal{N} = \mathcal{N}_0/\{0\}$ to denote the set of nodes excluding the substation node. See Fig. 1 for an example 5-bus network. Each node $i \in \mathcal{N}$ is associated with an active power injection p_i, a reactive power injection q_i, and a voltage magnitude v_i. We use p, q and v to denote the p_i, q_i, v_i stacked into a vector. We consider the linear distribution power flow model, known as the Simplified Distflow model, which is a linear approximation of the underlying nonlinear branch flow model in radial distribution networks [38] by neglecting the quadratic power loss terms. The Simplified Distflow model is a commonly used model in the voltage control literature, see e.g. [3], [5]–[9] and the references therein. In the Simplified Distflow model, p, q and v satisfy the following equations, $\forall j \in \mathcal{N}, i = \text{parent}(j)$,

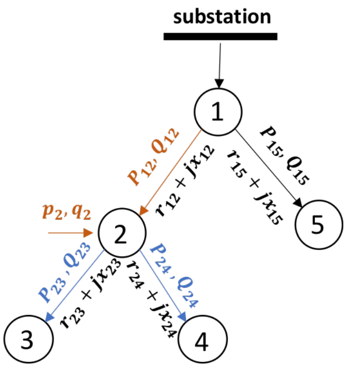

Fig. 1: A 5-bus radial distribution network.

$$-p_j = P_{ij} - \sum_{k:(j,k)\in\mathcal{E}} P_{jk}\,, \quad -q_j = Q_{ij} - \sum_{k:(j,k)\in\mathcal{E}} Q_{jk}\,, \tag{1a}$$

$$-v_j = v_i - 2(r_{ij}P_{ij} + x_{ij}Q_{ij})\,, (i,j) \in \mathcal{E} \tag{1b}$$

In Equation (1a), P_{ij} and Q_{ij} represent the active power and reactive power flow on line (i, j), and r_{ij} and x_{ij} are the line resistance and reactance. Briefly speaking, Equation (1a) presents the power conservation law at node j, e.g., in Figure 1, the power inflow at node 2 (orange) equals to the power outflow at node 2 (blue). Equation (1b) models the voltage drop from node i to node j. The above equation can also be rearranged and written in the vector form [3]:

$$\mathbf{v} = R\mathbf{p} + X\mathbf{q} + \mathbf{v}_0\mathbf{1} = X\mathbf{q} + \mathbf{v}^{env}. \tag{2}$$

Here we separate the voltage magnitude $\mathbf{v}$ into two parts: the controllable part $X\mathbf{q}$ that can be adjusted via adjusting reactive power injection $\mathbf{q}$ through various control devices, and the non-controllable part $\mathbf{v}^{env} = R\mathbf{p} + \mathbf{v}_0$ that is decided by the load and PV power $\mathbf{p}$. Matrix R, X are given as follows, $R_{ij} = 2\sum_{(h,k)\in P_i \cap P_j} r_{hk}\,, X_{ij} = 2\sum_{(h,k)\in P_i \cap P_j} x_{hk}$, where P_i is the set of lines on the unique path from the substation to bus i. Matrix X and R satisfy the following property, which is crucial for the stable control design.

Proposition 1. *Suppose $x_{ij}, r_{ij} > 0$ for all (i, j). Then, X and R are positive definite matrices.*

B. Voltage Control Problem Formulation

The goal of voltage control is to design control to lead the system voltage to reach the acceptable range $[\underline{v}, \overline{v}]$ under any system operating conditions at the lowest cost. Formally, voltage stability is defined as follows.

Definition 1 (Voltage stability)**.** *The closed loop system is stable if for any $\mathbf{v}^{env}$ and $\mathbf{v}(0)$, we have $\mathbf{v}(t)$ converges to the set $S_v = \{\mathbf{v} \in \mathbb{R}^n : \underline{v}_i \le v_i \le \bar{v}_i\}$ in the sense that*

$\lim_{t\to\infty}\mathrm{dist}(\mathbf{v}(t), S_v) = 0$ *and the distance is defined as* $\mathrm{dist}(\mathbf{v}(t), S_v) = \min_{\mathbf{v}'\in S_v} ||\mathbf{v}(t) - \mathbf{v}'||$.

Violations of the acceptable voltage range are often caused by a sudden change in the load or the generation, which can damage the appliances of the end users and even cause cascading failures if the system cannot return to the range promptly [28]. Therefore voltage stability is a vital requirement for the safe operation of the power systems.

With the requirement for voltage stability, the optimal voltage control problem can be formulated as follows,

$$\min_{\theta} \quad J(\theta) = \int_{t=0}^{\infty} \gamma^t \sum_{i=1}^{n} c_i(v_i(t), u_i(t))dt \tag{3a}$$

$$\text{s.t.} \quad \mathbf{v}(t) = X\mathbf{q}(t) + \mathbf{v}^{env} \tag{3b}$$

$$\dot{q}_i(t) = u_i(t) = g_{\theta_i}(v_i(t)) \tag{3c}$$

$$\text{Voltage stability holds.} \tag{3d}$$

The goal of the voltage control problem is to reduce the total cost (3a), which consists of two parts: the cost on voltage deviation and the cost of control actions. In particular, we consider $c_i(v_i(t), u_i(t)) = \eta_1[\max(v_i(t)-\bar{v}_i, 0)+\min(v_i(t)-\underline{v}_i, 0)]^2 + \eta_2(u_i(t))^2$. Here η_1, η_2 are coefficients that balance the cost of action with respect to the voltage deviation. We can set different η_1, η_2 at different nodes, and for simplicity, we choose the same η_1, η_2 across all nodes in the paper. Voltage dynamics of power system are represented by the power flow equation (3b). We envision that the reactive power control loop is embedded in an inverter control loop and operate at very fast timescales. Therefore, we use a continuous-time system to model the voltage/reactive power dynamics (unlike the more conventional discrete-time model). The control action u_i means the rate of change of the reactive power injection q_i in equation (3c), and we focus on the class of decentralized polices, $u_i(t) = g_{i,\theta_i}(v_i(t))$ only depends on local voltage measurement $v_i(t)$. Here θ_i is the policy parameter for the local policy g_{i,θ_i}, and $\theta = (\theta_i)_{i\in\mathcal{N}}$ is the collection of the local policy parameters and is also the decision variable in (3).

Transient cost vs. stationary cost. Our problem formulation is different from those in the literature [1]–[9] in the sense that the existing works typically consider the cost in stationarity, meaning the cost is evaluated at the fixed point or stationary point of the system. In contrast, our work considers the transient cost evaluated along the system trajectory, which is also an important metric for the performance of voltage control. An important future direction is to unify these two perspectives and design policies that can optimize both transient and stationary costs.

C. Existing Work: DDPG for Optimal Voltage Control

In order to solve the optimal voltage control problem in (3), one need the exact system dynamics, i.e., X. However, for distribution system, the exact network parameters are often unknown or hard to estimate in real systems [13]. Further, (3) is a nonlinear control problem Reinforcement learning provides a powerful paradigm for solving (3), by training a policy that maps the state to action via interacting with the environment, so as to minimize the loss function defined as (3a).

There are many RL algorithms to solve the policy minimization problem (3). In this paper, we focus on the class of RL algorithms called policy optimization. Generally speaking, the procedure is to run gradient methods on the policy parameter θ with step size η, $\theta \leftarrow \theta - \eta\nabla J(\theta)$. To approximate the gradient $\nabla J(\theta)$, one can use sampled trajectories such as REINFORCE [39] or value function approximation such as actor-critic methods [40]. As we are dealing with deterministic policies, one of the most popular choices is the Deep Deterministic Policy Gradient (DDPG) [29], where the policy gradient is approximated by

$$\nabla J(\theta) \approx \frac{1}{N}\sum_{i\in B} \nabla_u \hat{Q}_\phi(v, u)|_{v=v[i], u=g_\theta(v[i])} \nabla_\theta g_\theta(v)|_{v[i]}\,, \tag{4}$$

where $g_\theta(v)$ is the actor network, and $\{v[i], u[i]\}_{i\in B}$ are a batch of samples with batch size $|B| = N$ sampled from the replay buffer which stores historical state-action transition tuples. Here $\hat{Q}_\phi(v, u)$ is the value network (a.k.a critic network) that can be learned via temporal difference learning,

$$\min_{\phi} L(\phi) = E_{(v,u,r,v')}[Q_\phi(v, u) - (r + \gamma Q_\phi(v', g_\theta(v'))] \tag{5}$$

where v' is system voltage after taking action u and realization of v^{env}. For more details of DDPG, readers could refer to [29]. There have been a growing line of papers that use RL for voltage control [13], [21], [22]. In particular, [21] uses standard multi-agent DDPG for voltage control. However, in standard DDPG, stability is not an explicit requirement. It is more like an implicit regularization, because instability usually leads to high (or infinite) costs. Next, we will introduce our framework that guarantees stability in policy learning.

III. Main Results

We now introduce our main framework for stability constrained policy learning for voltage control. We start by stating our main result on the voltage control stability. We demonstrate the voltage stability constraint can be translated to a monotonicity constraint on the policy, which can be satisfied by smart design of monotone neural networks.

A. Key Idea: Monotonicity Guarantees Stability

As we mentioned in Section II-C, the lack of an explicit stability requirement in standard RL algorithms can lead to several issues. During the training phase, the policy may become unstable, causing the training process to terminate. Even after a policy is trained, there is no formal guarantee that the closed loop system is stable, which hinders the learned policy's deployment in real-world power systems where there is a very strong emphasis on stability. In order to explicitly constrain stability in policy learning, we constrain the search space of policy in a subset of stabilizing controllers from Lyapunov stability theory. Interestingly, we show that monotone policy guarantees stability for voltage control, which is presented in Theorem 1.

Theorem 1. *Suppose for all i, g_{i,θ_i} is a continuously differentiable function satisfying $g_{i,\theta_i}(v_i) = 0$ for $v_i \in [\underline{v}_i, \bar{v}_i]$. Further,* *suppose each g_{i,θ_i} is strictly monotonically decreasing on $(-\infty, \underline{v}_i]$ and $[\bar{v}_i, \infty)$, and satisfies $\lim_{v_i \to \infty} |g_{i,\theta_i}(v_i)| = \infty$. Then, the voltage stability defined in Definition 1 holds.*

Theorem 1 shows that the voltage stability condition in (3d) can be enforced by constraining the policy network to be monotone, which we will introduce in Section III-B. The key technique that underpins Theorem 1 is the Lyapunov stability theory, which involves defining a positive definite function $V(\cdot)$ that decreases along the system trajectory. Specially, we use Krasovskii's method [30] for constructing the Lyapunov function. For the voltage control problem with dynamics $\dot{\mathbf{v}} = f(\mathbf{v}, \mathbf{u}) = X\mathbf{u}$ and $\mathbf{u} = g_\theta(\mathbf{v}) = [g_{i,\theta_i}(v_i)]_{i \in \mathcal{N}}$, we consider the following Lyapunov function,

$$V(\mathbf{v}) = \frac{1}{2} f(\mathbf{v}, g_\theta(\mathbf{v}))^\top X^{-1} f(\mathbf{v}, g_\theta(\mathbf{v})) \tag{6}$$

where X is a positive definite matrix by Proposition 1. A sufficient condition for the system to be stable is that, the derivative of the Lyapunov function (6) satisfies the following condition,

$$\begin{aligned} \frac{d}{dt} V(\mathbf{v}(t)) &= (\nabla_{\mathbf{v}} V(\mathbf{v}))^\top \dot{\mathbf{v}} \\ &= \frac{1}{2} (X\mathbf{u})^\top \left[X^{-1} G(\mathbf{v}, \theta) + G(\mathbf{v}, \theta)^\top X^{-1} \right] \\ &\quad (X\mathbf{u}) < 0, \forall \mathbf{v} \notin \mathcal{S}_e \end{aligned} \tag{7}$$

where $G(\mathbf{v}, \theta) = \frac{\partial f(\mathbf{v}, \mathbf{u})}{\partial \mathbf{v}} + \frac{\partial f(\mathbf{v}, \mathbf{u})}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \mathbf{v}} = X \frac{\partial g_\theta(\mathbf{v})}{\partial \mathbf{v}}$ is the system Jacobian and $\mathcal{S}_e$ is the set of equilibrium points. This leads to our voltage stability condition,

$$\left[\frac{\partial g_\theta(\mathbf{v})}{\partial \mathbf{v}} + \frac{\partial g_\theta(\mathbf{v})}{\partial \mathbf{v}}^\top \right] \prec 0, \forall \mathbf{v} \notin \mathcal{S}_e \tag{8}$$

In particular, since the controller is decentralized where $u_i(t) = g_{i,\theta_i}(v_i(t))$ only depends on local voltage measurement $v_i(t)$, the Jacobian matrix $\frac{\partial g(\mathbf{v})}{\partial \mathbf{v}}$ is diagonal with the i-th element as $\frac{\partial g_{i,\theta_i}(v_i)}{\partial v_i}$. Therefore, conditions (8) can be met with each g_{i,θ_i} being strictly monotonically decreasing on $(-\infty, \underline{v}_i]$ and $[\bar{v}_i, \infty)$, and $g_{i,\theta_i}(v_i) = 0$ for $v_i \in [\underline{v}_i, \bar{v}_i]$. The detailed proof are as provided in Appendix A.

B. Algorithm Design

The proposed stability-constrained policy learning algorithm works as follows. At every training iteration k, we randomly generate an initial states $\{v_i(0)\}$ for all nodes $i = 1, ..., N$. Then we use the current policy $g^{(k)}_{i,\theta_i}(v_i(t))$ to generate a trajectory of length T, and store the (state, action, reward, next state) data pairs in a replay buffer. Next, we use random samples from the replay buffer to update the policy and value networks following Eq. (4) and Eq. (5). Specially, we parameterize the control policies $g_{i,\theta_i}(v_i), \forall i = 1, ..., N$ via monotone neural networks with a deadband in $[\underline{v}_i, \bar{v}_i]$. As shown in Theorem 1, such design guarantees voltage stability.

There are different approaches for monotone neural network architecture design in literature [37], [41], [42]. In this paper we follow the monotonic neural network design in [37, Lemma 3], which used a single hidden layer neural network with d hidden units and ReLU activation.

Corollary 1. *(Stacked ReLU Monotone Network [37, Lemma 3]) The stacked ReLU function constructed by Eq (9) is monotonic increasing for $x > 0$ and zero when $x \leq 0$.*

$$\xi^+(x; w^+, b^+) = (w^+)^\top \mathit{ReLU}(\mathbf{1}x + b^+) \tag{9a}$$

$$\text{where} \sum_{i=1}^{l} w_i^+ \geq 0, \forall l = 1, 2, ..., d \tag{9b}$$

$$b_1^+ = 0, b_l^+ \leq b_{l-1}^+, \forall l = 2, 3, ..., d \tag{9c}$$

The stacked ReLU function constructed by Eq (10) is monotonic increasing for $x < 0$ and zero when $x \geq 0$.

$$\xi^-(x; w^-, b^-) = (w^-)^\top \mathit{ReLU}(-\mathbf{1}x + b^-) \tag{10a}$$

$$\text{where} \sum_{i=1}^{l} w_i^- \leq 0, \forall l = 1, 2, ..., d \tag{10b}$$

$$b_1^- = 0, b_l^- \leq b_{l-1}^-, \forall l = 2, 3, ..., d \tag{10c}$$

Following Corollary 1, we parameterize the controller at bus i as $g_{i,\theta_i}(v_i) = -[\xi^+_{\theta_i}(v_i) + \xi^-_{\theta_i}(v_i)]$ where $\xi^+_{\theta_i}(v_i) : \mathbb{R} \to \mathbb{R}$ is monotonically increasing for $v_i > 0$ and zero when $v_i \leq 0$, and $\xi^-_{\theta_i}(v_i) : \mathbb{R} \to \mathbb{R}$ is monotonically increasing for $v_i < 0$ and zero otherwise. In addition, to incorporate the dead-band within range $v_i \in [\underline{v}_i, \bar{v}_i]$, we can simply set $w_1^+ = 0, b_2^+ = -\bar{v}_i$ and $w_1^- = 0, b_2^- = -\underline{v}_i$. We demonstrate the effectiveness of this approach using a case study in the next section.

IV. Case Study

We end the paper with a case study demonstrating the effectiveness of our approach for stability constrained policy learning for voltage control.

Experimental Setup Our evaluations focus on a Southern California Edison 56 bus distribution system with high penetration of photovoltaic (PV) generations. The detailed system parameters are given in [43]. Figure 2 provides the 56-bus distribution circuit, where there are 5 PV generators and controllers located at Buses 18, 21, 30, 45 and 53. Simulations of the power system dynamics use pandapower [44].

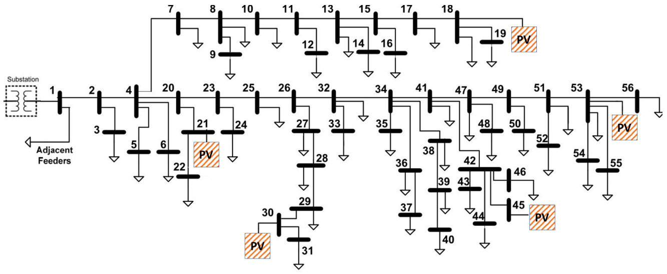

Fig. 2: Schematic diagram of SCE 56 bus distribution system with PV generations.

The nominal voltage magnitude at each bus is 12kV, and the acceptable range for operation is ±5% of the nominal value which is [11.4kV, 12.6kV]. We simulate three different scenarios: 1) High voltages: the PV generators are generating large amount of power, this corresponds to the day-time scenario in California where there is abundant sunshine that can result in high voltage issues at some buses. 2) Low voltages:

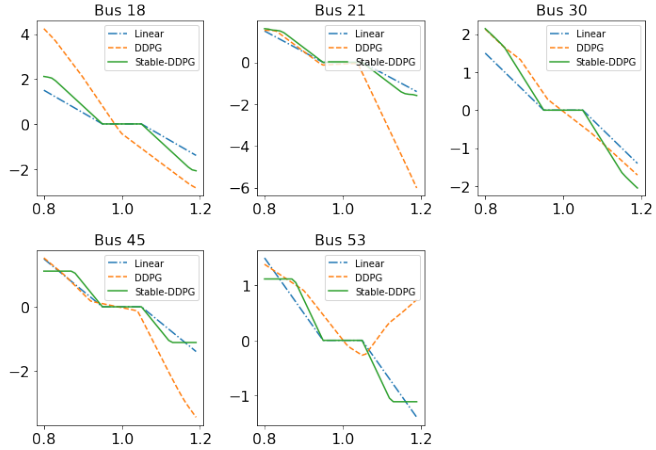

Fig. 3: Visualization of DDPG, Stable-DDPG and linear control policy at 5 PV buses. The x-axis is voltage (unit: kV) and the y-axis is control action (unit: MVar).

TABLE I: Performance of linear, DDPG and Stable-DDPG policies on 500 voltage violation scenarios. Note: reactive power consumption denotes the control cost.

	Voltage recovery time (steps)		Reactive power (MVar)	
Method	Mean	Std	Mean	Std
Linear	48.38	19.57	179.08	129.03
Stable-DDPG	**31.96**	14.49	**119.30**	89.05
DDPG	42.87	38.32	152.53	175.62

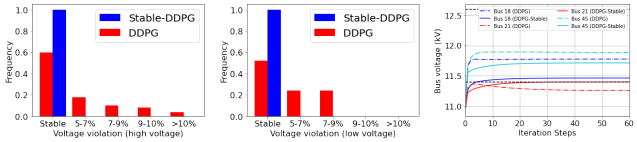

Fig. 4: Voltage stability of DDPG and Stable-DDPG.

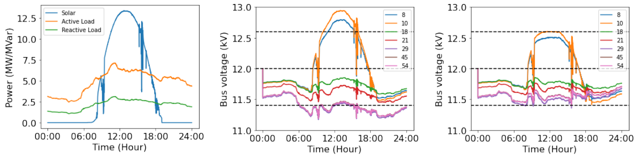

Fig. 5: Stable-DDPG test with real-world load and PV generation dataset. Left plot is the PV and aggregated load. Right two plots are the voltage without control and with Stable-DDPG.

the system is serving heavy loads without PV generation. It corresponds to a nighttime scenario when there is no sunshine but significant load, which results in low voltage issues at some buses. 3) A mix of high and low voltages: there are both high PV generation and heavy loads, which results in a mixture of high voltage issues at some buses and low voltages at others. For each scenario, we vary the PV output and the load to obtain different voltage conditions.

Our Approach and Baselines We incorporate the monotone policy network design with DDPG [29] to obtain the provable stable RL controller, and name our approach as Stable-DDPG. For baseline algorithms, we consider the following linear policy with deadband $d_i(v_i) = -[v_i - \overline{v}_i]^+ + [\underline{v}_i - v_i]^+$ (where $[x]^+ = \max(x, 0)$), which has been proposed and widely used in the power system control community [1], [3]. It guarantees stability but may lead to suboptimal control cost. We also compare the performance of Stable-DDPG against DDPG, which is suggested for voltage control in [21]. Standard DDPG minimizes the control cost without a formal stability guarantee. Details about the implementation and hyperparameters of Stable-DDPG and DDPG are provided in Appendix-A.

Experimental Results Figure 3 illustrates the control policy that is learned from Stable-DDPG and the baseline linear control [3]. Standard DDPG does not guarantee stability and thus can lead to "infinite" voltage recovery time and control cost. To obtain a reasonable comparison, we limit the max episode length to be $T = 100$, and compare the voltage recovery time (steps) and reactive power consumption (MVar) on 500 different voltage violation scenarios. Table I shows the results. In terms of control performance, the average voltage recovery time of Stable-DDPG is 31.96 steps, which saves about a third of the response time compared to the linear policy (48.38 steps). As the Stable-DDPG is able to drive back the voltage into normal operation state faster, the average transient control cost (computed as the sum of control cost before voltage stabilization) of Stable-DDPG is reduced by 33.38% compared to the linear policy. The standard deviation is high since the recovery time and control effort under different voltage violation conditions can be quite different. In addition, the lack of stability guarantee of standard DDPG can lead to higher control and state violation cost, which reflects in Table I that the control performance of standard DDPG is worse than the stable DDPG when averaging all scenarios.

Figure 4 compares the ability to achieve voltage stability of DDPG and Stable-DDPG under various test scenarios. The left plot shows the histogram of over-voltage ratio (i.e., $(v_T - v_0)^+/v_0$) and the middle plot shows the under-voltage ratio (i.e., $(v_0 - v_T)^+/v_0$). Our method achieves voltage stability in all scenarios, whereas DDPG may lead to voltage instability, with the final voltage beyond the ±5% range for 30-40% test scenarios. This is a serious issue since large voltage deviations violate regulatory requirements [27] and can cause cascading failures [28]. Figure 4 right shows a test case where DDPG leads to voltage instability at bus 21.

Finally, we test the proposed method using real-world data from DOE [8]. We simulate a massive solar penetration scenario where all buses are associated with PV and voltage controllers. We compare the voltage dynamics without voltage control and when Stable-DDPG is used. Corresponding to the time resolution of the PV and load trajectory, the proposed method adjust its control output every 6s. The voltage control results are given in Figure 5. There are severe voltage violations without control, due to the high volatility in load and PV generation. In contrast, Stable-DDPG quickly brings the voltage into the stable operation range, which further demonstrates its applicability in power system voltage control.

V. Conclusion

In this work, we propose a stability aware policy learning framework that formally guarantees the stability of RL in safety-critical systems. The key technique that underpins the proposed approach is to use Krasovskii's method for Lyapunov function construction and enforce the stability condition via monotone policy network design. We demonstrate the performance of the proposed method in real-world power system voltage control. Krasovskii's method is one way to construct the Lyapunov function, and a future direction is to incorporate other principled ways to construct Lyapunov functions in control theory.

References

[1] B. Zhang, A. D. Domínguez-García, and D. Tse, "A local control approach to voltage regulation in distribution networks," in *Proc. North American Power Symp.*, 2013, pp. 1–6.

[2] S. Bolognani, R. Carli, G. Cavraro, and S. Zampieri, "Distributed reactive power feedback control for voltage regulation and loss minimization," *IEEE Trans. Autom. Control*, vol. 60, no. 4, pp. 966–981, April 2015.

[3] N. Li, G. Qu, and M. Dahleh, "Real-time decentralized voltage control in distribution networks," in *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2014, pp. 582–588.

[4] H. Zhu and H. J. Liu, "Fast local voltage control under limited reactive power: Optimality and stability analysis," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3794–3803, 2016.

[5] G. Cavraro, S. Bolognani, R. Carli, and S. Zampieri, "The value of communication in the voltage regulation problem," in *Proc. IEEE Conf. on Decision and Control*, Las Vegas, NV, 2016.

[6] H. J. Liu, W. Shi, and H. Zhu, "Hybrid voltage control in distribution networks under limited communication rates," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 2416–2427, May 2019.

[7] Z. Tang, D. J. Hill, and T. Liu, "Fast distributed reactive power control for voltage regulation in distribution networks," *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 802–805, 2019.

[8] G. Qu and N. Li, "Optimal distributed feedback voltage control under limited reactive power," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 315–331, 2019.

[9] S. Magnússon, G. Qu, and N. Li, "Distributed optimal voltage control with asynchronous and delayed communication," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3469–3482, 2020.

[10] Y. Weng, Y. Liao, and R. Rajagopal, "Distributed energy resources topology identification via graphical modeling," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2682–2694, 2016.

[11] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, p. 484, 2016.

[12] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke *et al.*, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Conference on Robot Learning*. PMLR, 2018, pp. 651–673.

[13] Y. Chen, Y. Shi, and B. Zhang, "Data-driven optimal voltage regulation using input convex neural networks," *Electric Power Systems Research*, vol. 189, p. 106741, 2020.

[14] Y. Gao, W. Wang, and N. Yu, "Consensus multi-agent reinforcement learning for volt-var control in power distribution networks," *IEEE Trans. Smart Grid*, pp. 1–1, 2021.

[15] X. Sun and J. Qiu, "Two-stage volt/var control in active distribution networks with multi-agent deep reinforcement learning method," *IEEE Trans. Smart Grid*, pp. 1–1, 2021.

[16] Y. Zhang, X. Wang, J. Wang, and Y. Zhang, "Deep reinforcement learning based volt-var optimization in smart distribution systems," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 361–371, 2021.

[17] S. Mukherjee, R. Huang, Q. Huang, T. L. Vu, and T. Yin, "Scalable voltage control using structure-driven hierarchical deep reinforcement learning," *arXiv preprint arXiv:2102.00077*, 2021.

[18] P. Kou, D. Liang, C. Wang, Z. Wu, and L. Gao, "Safe deep reinforcement learning-based constrained optimal control scheme for active distribution networks," *Applied Energy*, vol. 264, p. 114772, 2020.

[19] H. Xu, A. D. Domínguez-García, and P. W. Sauer, "Optimal tap setting of voltage regulation transformers using batch reinforcement learning," *IEEE Trans. Power Syst.*, vol. 35, no. 3, pp. 1990–2001, May 2020.

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

[21] S. Wang, J. Duan, D. Shi, C. Xu, H. Li, R. Diao, and Z. Wang, "A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning," *IEEE Trans. Power Syst.*, 2020.

[22] W. Wang, N. Yu, Y. Gao, and J. Shi, "Safe off-policy deep reinforcement learning algorithm for Volt-VAR control in power distribution systems," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3008–3018, Jul. 2020.

[23] J. Duan, D. Shi, R. Diao, H. Li, Z. Wang, B. Zhang, D. Bian, and Z. Yi, "Deep-reinforcement-learning-based autonomous voltage control for power grid operations," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 814–817, Jan. 2020.

[24] D. Cao, W. Hu, J. Zhao, Q. Huang, Z. Chen, and F. Blaabjerg, "A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters," *IEEE Trans. Power Syst.*, 2020.

[25] H. Liu and W. Wu, "Two-stage deep reinforcement learning for inverter-based volt-var control in active distribution networks," *IEEE Trans. Smart Grid*, 2020.

[26] X. Chen, G. Qu, Y. Tang, S. Low, and N. Li, "Reinforcement learning for decision-making and control in power systems: Tutorial, review, and vision," *arXiv preprint arXiv:2102.01168*, 2021.

[27] "Global survey of regulatory approaches for power quality and reliability," Electric Power Research Institute, Palo Alto, CA, Tech. Rep., 2005.

[28] H. Haes Alhelou, M. E. Hamedani-Golshan, T. C. Njenda, and P. Siano, "A survey on power system blackout and cascading events: Research motivations and challenges," *Energies*, vol. 12, no. 4, p. 682, 2019.

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

[30] H. K. Khalil, *Nonlinear systems*, vol. 3.

[31] T. J. Perkins and A. G. Barto, "Lyapunov design for safe reinforcement learning," *Journal of Machine Learning Research*, vol. 3, no. Dec, pp. 803–832, 2002.

[32] Y.-C. Chang, N. Roohi, and S. Gao, "Neural lyapunov control," *Advances in neural information processing systems*, 2019.

[33] W. Jin, Z. Wang, Z. Yang, and S. Mou, "Neural certificates for safe control policies," *arXiv preprint arXiv:2006.08465*, 2020.

[34] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, "A lyapunov-based approach to safe reinforcement learning," *Advances in neural information processing systems*, 2018.

[35] S. M. Richards, F. Berkenkamp, and A. Krause, "The lyapunov neural network: Adaptive stability certification for safe learning of dynamical systems," in *Conference on Robot Learning*. PMLR, 2018, pp. 466–476.

[36] G. Manek and J. Z. Kolter, "Learning stable deep dynamics models," *Advances in neural information processing systems*, 2019.

[37] W. Cui and B. Zhang, "Reinforcement learning for optimal frequency control: A lyapunov approach," *arXiv preprint arXiv:2009.05654*, 2020.

[38] M. E. Baran and F. F. Wu, "Network reconfiguration in distribution systems for loss reduction and load balancing," *IEEE Power Engineering Review*, vol. 9, no. 4, pp. 101–102, 1989.

[39] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[40] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Advances in neural information processing systems*. Citeseer, 2000, pp. 1008–1014.

[41] J. Sill, "Monotonic networks," in *Proceedings of the 10th International Conference on Neural Information Processing Systems*, 1997, pp. 661–667.

[42] H. Daniels and M. Velikova, "Monotone and partially monotone neural networks," *IEEE Transactions on Neural Networks*, vol. 21, no. 6, pp. 906–917, 2010.

[43] M. Farivar, R. Neal, C. Clarke, and S. Low, "Optimal inverter var control in distribution systems with high pv penetration," in *2012 IEEE Power and Energy Society general meeting*. IEEE, 2012, pp. 1–7.

[44] L. Thurner, A. Scheidler, J. Dollichon, F. Sch?fer, J.-H. Menke, F. Meier, S. Meinecke *et al.*, "pandapower - convenient power system modelling and analysis based on pypower and pandas," University of Kassel and Fraunhofer Institute for Wind Energy and Energy System Technology, Tech. Rep., 2016. [Online]. Available: http://pandapower.readthedocs.io/en/v1.2.2/_downloads/pandapower.pdf

[45] J.-J. E. Slotine, W. Li *et al.*, *Applied nonlinear control*. Prentice hall Englewood Cliffs, NJ, 1991, vol. 199, no. 1.

Appendix

Appendix-A: Proof of the Main Results

For the proofs of Theorem 1, we use a generalization of Lyapunov's direct method, known as LaSalle's Invariance Principle. We provide a version of it below, adapted from [45], which we slightly change to fit the rest of the paper.

Proposition 2 (Theorem 3.4, 3.5 in [45])**.** *For dynamical system $\dot{x} = F(x)$, suppose $V : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function satisfying $V(x) \geq 0$, $\dot{V}(x) = [\nabla_x V(x)]^\top F(x) \leq 0, \forall x \in \mathbb{R}^n$. Let $S_e = \{x : \dot{V}(x) = 0\}$. If there exists $a \in \mathbb{R}$ such that the level set $L_a := \{x : V(x) \leq a\}$ is bounded, then for any $x(0) \in L_a$ we have $\mathrm{dist}(x(t), S_e) \to 0$. Further, if V is radially unbounded, i.e. $V(x) \to \infty$ as $\|x\| \to \infty$, then, for any $x(0) \in \mathbb{R}^n$, we have $\mathrm{dist}(x(t), S_e) \to 0$.*

With the LaSalle's Invariance Principle, we are now ready to prove Theorem 1.

Proof of Theorem 1. For the voltage control problem (3), let $\mathbf{v}(t)$ be the state, $\mathbf{u}(t)$ be the action. Then, we have $\dot{\mathbf{v}} = X\dot{\mathbf{q}} = X\mathbf{u} = Xg_\theta(\mathbf{v})$, where $g_\theta(\mathbf{v}) = [g_{i,\theta_i}(v_i)]_{i\in\mathcal{N}}$ are the decentralized policies. We consider the following Lyapunov function,

$$V(\mathbf{v}) = \frac{1}{2} g_\theta(\mathbf{v})^\top X g_\theta(\mathbf{v}).$$

Clearly, V is positive definite and is radially unbounded by the assumptions of Theorem 1. By LaSalle's Invariance principle, to prove Theorem 1, we only need to show the following claim.

Claim: $\frac{d}{dt}V(\mathbf{v}(t)) \leq 0$, and $\frac{d}{dt}V(\mathbf{v}(t)) = 0$ only when $\mathbf{v} \in S_v$, where we recall $S_v = \{\mathbf{v} \in \mathbb{R}^n : \underline{v}_i \leq v_i \leq \bar{v}_i\}$.

It is easy to check that,

$$\frac{d}{dt}V(\mathbf{v}(t)) = (\nabla_{\mathbf{v}} V(\mathbf{v}))^\top \dot{\mathbf{v}} = (Xg_\theta(\mathbf{v}))^\top \frac{\partial g_\theta(\mathbf{v})}{\partial \mathbf{v}} (Xg_\theta(\mathbf{v})).$$

Note that $Xg_\theta(\mathbf{v}) = 0$ if and only if $\mathbf{v} \in S_v$. Therefore, to show the claim, it suffices to show the following two conditions,

$$\frac{\partial g_\theta(\mathbf{v})}{\partial \mathbf{v}} \preceq 0, \forall \mathbf{v} \in \mathbb{R}^n, \tag{11a}$$

$$Xg_\theta(\mathbf{v}) \notin \ker(\frac{\partial g_\theta(\mathbf{v})}{\partial \mathbf{v}})/\{0\}, \forall \mathbf{v} \in \mathbb{R}^n, \tag{11b}$$

where $\ker(\cdot)$ denotes the null space of a matrix.

We first check (11a). Note $\frac{\partial g_\theta(\mathbf{v})}{\partial \mathbf{v}}$ is a diagonal matrix whose i'th entry is $g'_{i,\theta_i}(v_i)$, which is nonnegative as $g_{i,\theta_i}(v_i)$ is monotonically decreasing. Therefore, (11a) is true. To check (11b), suppose $\mathbf{v}$ is such that,

$$0 = \frac{\partial g_\theta(\mathbf{v})}{\partial \mathbf{v}} X g_\theta(\mathbf{v}). \tag{12}$$

Then, let $C \subset \mathcal{N}$ be the set of indices $\{i \in \mathcal{N} : v_i \in [\underline{v}_i, \bar{v}_i]\}$, and $C' = \mathcal{N}/C$. In the following, we will show that $C' = \emptyset$. Suppose the contrary is true, i.e. $C' \neq \emptyset$. Note for $i \in C$, $g_{i,\theta_i}(v_i) = 0$, and further, the i'th diagonal entry of $\frac{\partial g_\theta(\mathbf{v})}{\partial \mathbf{v}}$ is zero. For $i \in C'$, the i'th diagonal entry of $\frac{\partial g_\theta(\mathbf{v})}{\partial \mathbf{v}}$ is non-zero. These show that (12) is equivalent to

$$X_{C',C'} g_{C',\theta_{C'}}(\mathbf{v}_{C'}) = 0, \tag{13}$$

where $X_{C',C'}$ is the submatrix of X corresponding to the rows and columns in C', and $g_{C',\theta_{C'}}(\mathbf{v}_{C'})$ is the subvector of $g_\theta(\mathbf{v})$ corresponding to the indices in C'. Since $X_{C',C'}$ is positive definite as it is a principal submatrix of X, we have (13) indicates that, for any $i \in C'$, $g_{i,\theta_i}(v_i) = 0 \Rightarrow v_i \in [\underline{v}_i, \bar{v}_i]$. Per the definition of C, this shows that such i must be an element of C, which is impossible since C' and C are disjoint. Therefore, we have a contradiction, and as a result, $C' = \emptyset$ and $g_{i,\theta_i}(v_i) = 0, \forall i \in \mathcal{N}$. Therefore, we have $Xg_\theta(\mathbf{v}) = 0$, and hence $Xg_\theta(\mathbf{v}) \notin \ker(\frac{\partial g_\theta(\mathbf{v})}{\partial \mathbf{v}})/\{0\}$. As a result, (11b) holds. Therefore, the claim is proven and the proof of Theorem 1 follows. □

Appendix-B: Simulation Details

We use Pytorch to build all RL models and run the training process in MacBook Pro Personal Laptop with 16 GB 2400 MHz DDR4 memory and 2.2 GHz Intel Core i7 processor. The reward function for training is $c(\mathbf{v}, \mathbf{u}) = 100(\mathrm{dist}(\mathbf{v}, S_v))^2 + 50||\mathbf{u}||_2^2$, where $\mathbf{v}$ is voltage vector and $\mathbf{u}$ is the policy output. Table II shows the hyperparameters used for both DDPG and Stable-DDPG. For policy network design, Stable-DDPG requires the policy network to be monotone, and the monotone network architecture based on [37] only applies to single layer neural network. Thus, we use a one fully-connected layer neural network for Stable-DDPG and a larger capacity model (two-layer fully-connected neural network) for standard DDPG. As shown in Fig 6, Stable-DDPG has higher initial reward and lower variance because of the monotone structure (i.e., stability guarantee by design).

TABLE II: Hyperparameters for DDPG and Stable-DDPG

Hyper-parameters	DDPG	Stable-DDPG
Policy Net	100-100	100
Q function net	100-100	100-100
Discount factor (γ)	0.99	0.99
Policy net learning rate	1e-4	1e-4
Q function net learning rate	2e-4	2e-4
Action noise	Gaussian(0, 0.05)	Gaussian(0, 0.05)
Maximum replay buffer size	1000000	1000000
Target network update ratio	1e-2	1e-2
Batch size	256	256
Activation function	ReLU	ReLU
Training episode	600	200
Episode length	30	30
State dimension (v_i)	1	1
Action dimension (u_i)	1	1

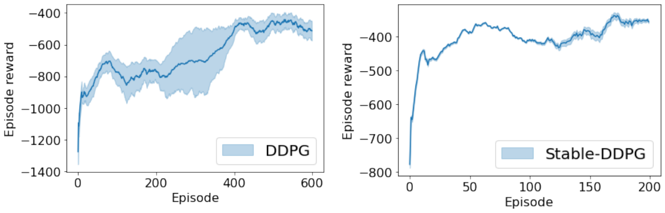

Fig. 6: Comparison of training episodic reward from DDPG and Stable-DDPG. The mean and standard deviations are evaluated based on 5 random seeds.