Reinforcement Learning-based Smart Inverter Control with Polar Action Space in Power Distribution Systems

Farzana Kabir, *Student Member, IEEE,* Yuanqi Gao, *Member, IEEE,* and Nanpeng Yu, *Senior Member, IEEE,*

***Abstract*— To tackle the challenge of voltage regulation under high solar photovoltaics (PV) penetration, the slow timescale control of conventional voltage regulating devices can be combined with fast timescale control of smart inverters. In this paper, we develop a two-timescale Volt-VAR control (VVC) framework. The slow timescale control of voltage regulating devices is achieved by a model-based approach. The fast timescale control of smart inverters is attained with a reinforcement learning-based method. The deep deterministic policy gradient (DDPG) algorithm is adopted to control the setpoints of both real and reactive power of smart inverters. The control policy of smart inverters is learned from the historical operational data without relying on accurate distribution network secondary circuit parameters. Simulation results on the IEEE 34-bus feeder show that the proposed framework can determine near optimal set points of smart inverters in real-time operations. Compared with existing reinforcement learning based smart inverter control, our approach achieves lower line losses, voltage deviations, and active power curtailment.**

***Index Terms*— Deep deterministic policy gradient, Volt-VAR control, smart inverters, reinforcement learning, high PV penetration, two timescale**

I. Introduction

Solar photovoltaics (PV) is projected to constitute 46% of total renewable generation by 2050, increasing from only 13% in 2018 [1] due to a rapid drop in cost [2]. However, solar energy is highly intermittent due to cloud cover and shading, fluctuating up to 15% of their nameplate ratings within one-minute intervals [3]. The increasing solar PV penetration in power distribution networks poses serious operational challenges, particularly in maintaining an appropriate feeder-wide voltage profile.

To keep the feeder voltage profile in a reasonable range, conventional Volt-VAR control (VVC) determines the optimal hourly set points for voltage regulating devices such as voltage regulators, on-load tap changers (OLTCs), and capacitor banks. However, these voltage regulating devices are slow operating mechanical equipment and are insufficient to adapt to distribution systems with fast and significant voltage fluctuations due to high solar PV penetration. Chronic voltage fluctuations can also lead to frequent operations of voltage regulating devices which will shorten their life cycles and increase maintenance costs [4].

To mitigate frequent voltage variations in distribution feeders with high solar PV penetration, smart solar PV inverter based VVC has been studied. Smart inverters provide fast and continuous active and reactive power control with low operational costs. Besides, they support two-way communications, which allow remote control systems to change inverter setpoints. This opens up considerable opportunities for utilities to integrate distributed solar PV systems into the VVC framework. The IEEE 1547a-2020 standard allows smart inverters to participate in grid voltage regulation [5].

Previous studies on inverter control consider varying the reactive power generation of solar PV systems using centralized [6]–[8], distributed [9]–[13], or local control approaches [14]–[16]. Centralized and distributed control solve an optimal power flow (OPF) problem to determine the inverter reactive power generation. Local control approaches calculate the reactive power generation using droop control. However, controlling only reactive power may yield low feeder power factors and cause high network currents. In fact, smart inverters can also curtail solar PV systems' active power generation to regulate feeder voltage [17], [18]. An optimization-based centralized approach is developed to determine both active and reactive power setpoints for smart inverters of solar PV systems in [19].

Recently, researchers have been developing two-timescale model-based VVC by supplementing the conventional slow timescale VVC with fast timescale smart inverter control [20]–[23]. However, the decision variables of the fast timescale smart inverter control only include reactive power setpoints. References [21] and [22] formulate the VVC as an OPF problem and propose to solve it using centralized optimization. The controllable devices on the slow timescale include capacitor banks [21], [22] and OLTCs [22], [23].

The model-based optimization approaches [24] rely on accurate and complete distribution network topology [25], [26] and parameter information [27]. However, the secondary feeders' phase connection information is usually not accurate [28]. To address these problems, researchers have developed data-driven control approaches for slow timescale VVC problems [8], [29]–[31] and fast timescale smart inverter control problem [32] using reinforcement learning (RL) algorithms. A two-timescale VVC framework considering both slow timescale voltage regulating devices and fast timescale smart inverter control is developed in [33]. For the slow timescale, deep Q-learning is used to determine the switching schedule of capacitors. For the fast timescale, an optimization-based approach is adopted to control the smart inverters. Many existing data-driven approaches need accurate line parameters and power injections at every bus which might not be available in real-time operations [34].

There are two main drawbacks of the existing data-driven VVC framework involving smart inverters. First, the existing

F. Kabir, Y. Gao, and N. Yu are with the Department of Electrical Engineering, University of California, Riverside, Riverside, CA, 92521-0429 USA e-mail: (fkabi001@ucr.edu, ygao024@ucr.edu, nyu@ece.ucr.edu).

approaches only consider changing the reactive power setpoints of smart inverters [32] and ignore the fact that active power could be curtailed for solar PV systems during certain circumstances. Second, the primary feeders' model is much more reliable than that of the secondary feeders. Thus, in the two-timescale VVC framework, the fast timescale control involving smart inverters in the secondary feeders should be data-driven and the slow timescale control involving the primary feeder can be handled with a model-based approach.

In this paper, we fill the knowledge gap by developing a two-timescale data-driven Volt-VAR control method, which does not rely on secondary feeder information. Note that our method still requires knowledge of the primary feeder, which is often readily available. Furthermore, we design a polar action space set up to jointly determine the active and reactive power setpoints of smart inverters. Specifically, on the slow timescale, a centralized optimization-based approach is adopted to determine the tap positions of voltage regulators, OLTCs, and switchable capacitor banks. On the fast timescale, a deep deterministic policy gradient (DDPG)-based algorithm is employed to determine the set points of real and reactive power of smart inverters.

The unique contributions of this paper are summarized below.

- We develop a reinforcement learning-based two-timescale VVC for distribution networks without requiring secondary feeders' topology or parameter information.
- We design a polar action space for reinforcement learning-based smart inverter control. This design allows joint determination of real and reactive power setpoints while explicitly enforcing maximum power capability constraint.
- The degradation costs of the smart inverters are carefully modeled in the sequential decision-making process of the VVC problem.

The rest of the paper is organized as follows. Section II presents an overview of the two-timescale VVC problem. Section III discusses the problem formulation of the slow timescale VVC and fast timescale smart inverter control. Section IV presents the proposed two-timescale VVC algorithms. Section V shows the numerical study results. Finally, Section VI states the conclusions.

II. Two-timescale VVC Framework

We consider a power distribution system with both conventional voltage regulating devices and smart inverters. The smart inverters control the real and reactive power setpoints of solar PV systems. A generic power distribution network being modeled is shown in Fig. 1. The overall framework of the two-timescale VVC is shown in Fig. 2. In the slow timescale VVC, the optimal tap positions and switching schedules of the voltage regulator, OLTCs, and capacitor banks are determined using a centralized optimization-based method on an hourly basis τ. Within each hour, the tap and switching positions of these voltage regulating devices are kept fixed. The technical method of the slow timescale VVC is discussed in detail in Subsection III-B. In the fast timescale VVC, the real and reactive power setpoints of smart inverters

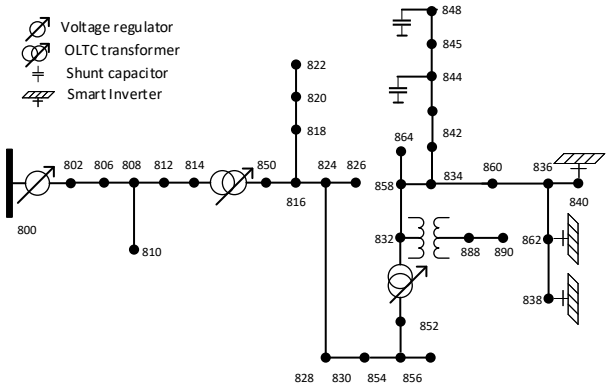

Fig. 1: Diagram of a typical power distribution network with voltage regulating devices and smart inverters.

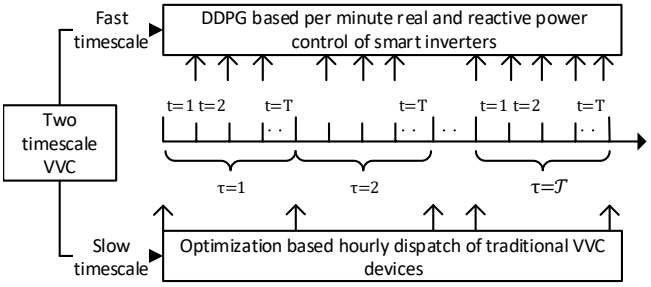

Fig. 2: The overall framework for two-timescale VVC

are determined every minute t to mitigate voltage violations caused by rapid fluctuations in the maximum potential output of solar PV systems. The smart inverter dispatch schedule is determined by the deep deterministic policy gradient algorithm which does not rely on an accurate secondary feeder model. The fast timescale VVC problem using smart inverters is formulated in Section III-C.

III. Problem Formulation

A. Problem Setup

Let us consider a radial distribution feeder of N buses represented by $\mathcal{G} := (\mathcal{N}, \mathcal{L})$. Here, $\mathcal{N} := \{1, \ldots, N\}$ is the collection of all nodes and $\mathcal{L} := \{(m, n) \subset \mathcal{N} \times \mathcal{N}\}$ is the collection of edges representing distribution line segments. Let r_{ij} and x_{ij} be resistance and reactance of the distribution line between node i and j. We assume that the distribution network is relatively balanced.

Let v_i denote the complex voltage phasor at node i for $i \in \mathcal{N}$ and u_i denote the square of the corresponding voltage magnitude. Let I_{ij} be the complex current flowing from node i to node j and ℓ_{ij} be the square of the corresponding current magnitude. Let P_{ij} and Q_{ij} be the real and reactive power flowing over the line connecting nodes i and j. Let p_i^g and q_i^g be the real and reactive power generation from the smart inverter connected solar PV system at node i, and p_i^c and q_i^c be the real and reactive power demand at node i. Let p_i^G and q_i^G be the total real and reactive power generation respectively at node i from the smart inverter connected solar

PV systems and switchable capacitors. Let $p_i + jq_i$ be the net complex power injection at node i where $p_i := p_i^G - p_i^c$ and $q_i := q_i^G - q_i^c$. Let $\bar{p}_{it}^g$ be the available solar PV production at time t for smart inverter i, which is determined by solar irradiance and the smart inverters' nameplate capacity $\bar{S}_i$. At any time t, the real and reactive power generation from smart inverters, electric demand $p_{it}^g, q_{it}^g, p_{it}^c, q_{it}^c$, and the settings of voltage regulators, OLTCs, and capacitor banks determine the voltages and power flows on the distribution network. The problem formulation of the slow timescale VVC using voltage regulating devices and fast timescale VVC using smart inverters are presented in the following two subsections.

B. Slow Timescale VVC Using Voltage Regulation Devices

For the slow timescale VVC subtask, the controllable devices include voltage regulators, OLTCs, and switchable capacitor banks. Voltage regulators are typically placed at the reference bus. Each of the voltage regulators and OLTCs has K tap positions with a step size of C^{reg} and C^{tsf} corresponding to the change in turns ratios. The series and shunt impedance of the voltage regulating devices can be neglected since their values are very small. Switchable capacitor banks are installed at different locations on the feeder to provide local voltage support. Let q_i^{cap} be the reactive power generation from the capacitor bank. Let tap_τ^{reg} and tap_τ^{tsf}, and tap_τ^{cap} indicate the tap position of the voltage regulators, OLTCs, and the switch status of the capacitor banks respectively at time τ.

For the slow timescale VVC subtask, the reactive power setpoints of smart inverters are assumed to be 0. Thus, $q_{j\tau}^G = q_{j\tau}^{cap}$ at every node j. The objective of the slow timescale VVC is to minimize the sum of line loss $C_e r_{ij} \ell_{ij\tau}$ and voltage deviation cost $C_V \left(u_{i\tau} - 1\right)^2$ at the beginning of each hour τ, where C_v and C_e are voltage deviation cost (\$/$volt$) and electricity price (\$/MWh) respectively. The slow timescale VVC is formulated as a mixed-integer nonlinear programming (MINLP) problem as follows:

$$\min_{\boldsymbol{X}} \sum_{(i,j)\in\mathcal{L}} C_e r_{ij} \ell_{ij\tau} + \sum_{i\in\mathcal{N}} C_v \left(u_{i\tau} - 1\right)^2 \tag{1}$$

$$\text{s.t. } P_{ij\tau} = \sum_{k:(j,k)\in\mathcal{L}} P_{jk\tau} + r_{ij}\ell_{ij\tau} + p_{j\tau}^c - \bar{p}_{j\tau}^g \quad \forall (i,j) \in \mathcal{L} \tag{2}$$

$$Q_{ij\tau} = \sum_{k:(j,k)\in\mathcal{L}} Q_{jk\tau} + x_{ij}\ell_{ij\tau} + q_{j\tau}^c - q_{i\tau}^{cap} \quad \forall (i,j) \in \mathcal{L} \tag{3}$$

$$u_{j\tau}/a_{ij\tau}^2 = u_{i\tau} - 2\left(r_{ij}P_{ij\tau} + x_{ij}Q_{ij\tau}\right) + \left(r_{ij}^2 + x_{ij}^2\right)\ell_{ij\tau} \quad \forall (i,j) \in \mathcal{L} \tag{4}$$

$$u_{1\tau} = \left(u^{ref} + tap_\tau^{reg} \times C^{reg}\right) \tag{5}$$

$$\ell_{ij\tau} = \frac{P_{ij\tau}^2 + Q_{ij\tau}^2}{u_{i\tau}} \quad \forall (i,j) \in \mathcal{L} \tag{6}$$

$$\boldsymbol{X} := \left(\boldsymbol{P}_\tau, \boldsymbol{Q}_\tau, \boldsymbol{u}_\tau, \boldsymbol{\ell}_\tau, \mathbf{tap}_\tau^{reg}, \mathbf{tap}_\tau^{tsf}, \mathbf{tap}_\tau^{cap}\right) \tag{7}$$

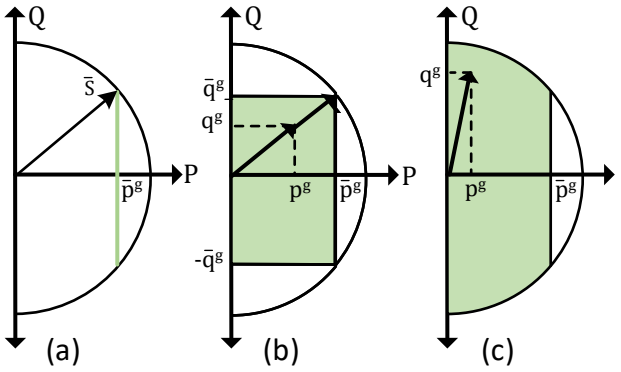

Fig. 3: (a) Action space for reactive power control strategy, (b) Rectangular action space for inverter control with both real and reactive power, (c) Polar action space for inverter control with both real and reactive power.

where $a_{ij\tau} = 1 + tap_\tau^{tsf} \times C^{tsf}$ if there is an OLTC on branch (i,j) and $a_{ij\tau} = 1$ otherwise. Note that MINLP is not as scalable as the state-of-the-art MILP solvers. As a result, formulating the slow timescale VVC as a MINLP problem might not be a feasible approach as the size of the distribution network grows and the number of discrete tap positions of the conventional VVC devices increases.Also note that in our current formulation, the slow timescale VVC controller is not aware of the reactive power from the smart inverter in the coming hour. However, we shall see in our numerical study section that, this formulation still improve the baseline without leading to the canceling effect of the two controllers in different timescales.

C. Fast Timescale VVC by Smart Inverters

1) Smart Inverter Control Strategies: A solar PV inverter has a maximum apparent power capability $\bar{S}_i > \max\left(\bar{p}_{it}^g\right)$. There are three operational strategies for smart solar PV inverters:

a) Reactive Power Control Strategy: If the i-th solar PV inverter allows reactive power control only, then the set of its operating points F_i^{RPC} is defined as:

$$F_i^{RPC} := \left\{ \left(p_{it}^g, q_{it}^g\right) \middle| p_{it}^g = \bar{p}_{it}^g, |q_{it}^g| \le \sqrt{\bar{S}_i^2 - (\bar{p}_{it}^g)^2} \right\}$$

Under this control strategy, the active power output is the available solar PV generation, and the reactive power output is limited by the inverter rating. The set F_i^{RPC} is represented by the vertical line segment in Fig. 3(a). If the inverter is not oversized, then smart inverter can not provide reactive power compensation when $\bar{p}_{it}^g = \bar{S}_i$. With oversized inverters, the entire inverter rating can be utilized to supply reactive power when no active power is produced.

b) Real and Reactive Power Control with Rectangular Operating Space: Under this control strategy, smart inverters are allowed to adjust both active and reactive power. However, the reactive power compensation is limited by the inverter rating and available solar PV production $\bar{p}_{it}^g$ at time t with $\bar{q}_{it}^{gR} = \sqrt{\bar{S}_i^2 - (\bar{p}_{it}^g)^2}$. Thus, the smart inverter operating

space is a rectangle as shown in Fig. 3(b). The set of possible operating points of the smart inverter is given by:

$$F_i^{RPCR} := \left\{ (p_{it}^g, q_{it}^g) \,\middle|\, 0 \le p_{it}^g \le \bar{p}_{it}^g, |q_{it}^g| \le \bar{q}_{it}^{gR} \right\}$$

Under this control strategy, when $0 \le p_{it}^g < \bar{p}_{it}^g$, active power curtailment takes place. The amount of real power curtailment p_{it}^C equals $\bar{p}_{it}^g - p_{it}^g$.

c) Real and Reactive Power Control with Polar Operating Space: Under this strategy, solar PV inverters are allowed to adjust both active and reactive powers. The reactive power compensation is limited by the inverter rating and the actual solar PV production p_{it}^g at time t as in $\bar{q}_{it}^{gP} = \sqrt{\bar{S}_i^2 - (p_{it}^g)^2}$, which makes the inverter operating space a curtailed semi-circle as shown in Fig. 3(c). Here, $\bar{q}_{it}^{gP} > \bar{q}_{it}^{gR}$ when active power curtailment take place, i.e. $p_{it}^g < \bar{p}_{it}^g$. Consequently, the set of possible operating points is given by

$$F_i^{RPCP} := \left\{ (p_{it}^g, q_{it}^g) \,\middle|\, 0 \le p_{it}^g \le \bar{p}_{it}^g,\ |q_{it}^g| \le \bar{q}_{it}^{gP} \right\}$$

2) Optimization based Fast Timescale Inverter Control: If the power distribution network model is complete and accurate, the optimal setpoints of smart inverters can be found by solving the following optimization problem at every time slot t within each interval τ. The tap positions of the voltage regulator, OLTC transformers, and capacitor banks are available from the last interval on the slow timescale. In addition to minimizing line loss and voltage deviation, the active power curtailment cost $C_c \left| \bar{p}_{it}^g - p_{it}^g \right|$ of each inverter is minimized where C_c is the active power curtailment cost $\$/MWh$. By relaxing the nonconvex quadratic equality constraint (13), the optimization problem can be converted to a Second Order Cone Program (SOCP) defined over a convex feasible set [21]. It has been shown that SOCP relaxation can be exact under certain conditions in the sense that the equality in (13) holds for optimal solutions [35], [36]. However, reverse power flows from extreme solar PV generation and an objective minimizing voltage deviation can lead to non-zero duality gap and non-physical solutions [37]–[39]. Nevertheless, our numerical study shows that the non-zero duality gap issue does not occur, which confirms the global optimality and exactness of the optimization-based benchmark.

$$\min_{\boldsymbol{X}} \sum_{(i,j)\in\mathcal{L}} C_e r_{ij} \ell_{ijt} + \sum_{i\in\mathcal{N}} \left[C_v \left(u_{it} - 1 \right)^2 + C_c \left| \bar{p}_{it}^g - p_{it}^g \right| \right] \tag{8}$$

$$\text{s.t} \quad P_{ijt} = \sum_{k:(j,k)\in\mathcal{L}} P_{jkt} + r_{ij}\ell_{ijt} + p_{jt}^c - p_{jt}^g \ \forall (i,j) \in \mathcal{L} \tag{9}$$

$$Q_{ijt} = \sum_{k:(j,k)\in\mathcal{L}} Q_{jkt} + x_{ij}\ell_{ijt} + q_{jt}^c - q_{jt}^g - q_{j\tau}^{cap}$$
$$\forall (i,j) \in \mathcal{L} \tag{10}$$

$$u_{jt}/a_{ij\tau}^2 = u_{it} - 2\left(r_{ij}P_{ijt} + x_{ij}Q_{ijt}\right) + \left(r_{ij}^2 + x_{ij}^2\right)\ell_{ijt}$$
$$\forall (i,j) \in \mathcal{L} \tag{11}$$

$$u_{1t} = \left(u^{ref} + tap_\tau^{reg} \times C^{reg} \right) \tag{12}$$

$$\ell_{ijt} = \left(P_{ijt}^2 + Q_{ijt}^2 \right) / u_{it} \quad \forall (i,j) \in \mathcal{L} \tag{13}$$

$$0 \le p_{it}^g \le \bar{p}_{it}^g, -\bar{q}_{it}^g \le q_{it}^g \le \bar{q}_{it}^{gP} \quad \forall i \in \mathcal{N} \tag{14}$$

$$\boldsymbol{X} := (\boldsymbol{P}_t, \boldsymbol{Q}_t, \boldsymbol{p}_t^g, \boldsymbol{q}_t^g, \boldsymbol{u}_t, \boldsymbol{\ell}_t) \tag{15}$$

Note that the tap position variables $a_{ij\tau}$, $q_{j\tau}^{cap}$, and tap_τ^{reg} are taken from the last interval of the slow timescale VVC. In the future, we plan to further enhance the model prediction control (MPC)-based fast timescale VVC algorithm by taking the inverter degradation into account. This makes the baseline algorithm more consistent with the proposed reinforcement learning-based algorithm.

IV. Two-timescale VVC using DDPG

A. Fast Timescale VVC as a Markov Decision Process

We briefly review the basics of the Markov decision process (MDP). An MDP can be defined as a tuple consists of a state space $\mathcal{S}$, an action space $\mathcal{A} = \Re^M$ (M is the dimension of the action space), an initial state distribution $p(s_1)$, a transition probability $p(s_{t+1}|s_t, a_t)$, and a reward function $R : \mathcal{S} \times \mathcal{A} \in \Re$. The agent interacts with the environment $\mathcal{E}$ according to some policy $\mu : \mathcal{S} \to A$ to generate trajectories of the form $s_1, a_1, r_1, \ldots, s_t, a_t, r_t, \ldots, s_T, a_T, r_T$, where $r_t = R(s_t, a_t)$. The return from a state is defined as the sum of discounted future reward $G_t = \sum_{i=t}^{T} \gamma^{(i-t)} R(s_i, a_i)$ with a discounting factor $\gamma \in [0, 1]$. The goal is to learn a policy which maximizes the expected return from the initial state $J = \mathbb{E}_{s\sim p(s_1)} \mathbb{E}_\mu [G_t | s_1 = s]$

To formulate the fast timescale VVC problem as an MDP, the distribution system operator or controller is treated as the agent and the distribution network is treated as the environment. We define the state, action, and reward function as follows:

a) State: The state consists of real and reactive power injection of inverters $\boldsymbol{p}_t^g, \boldsymbol{q}_t^g$, and loads $\boldsymbol{p}_t^c, \boldsymbol{q}_t^c$ at relevant nodes at time t, solar PV production potential of the inverters determined by solar irradiance and technical parameters of the respective PV systems $\bar{\boldsymbol{p}}_t^g$, voltage magnitude at each bus $|\boldsymbol{v}_t|$, and current tap positions of voltage regulating devices $tap^{reg}, tap^{tsf}, tap^{cap}$.

b) Action: In the VVC strategy adopted in this paper, the smart inverters are allowed to adjust both active and reactive power outputs. The active power provided by the smart inverter i can be expressed by $p_{it}^g = a_p \bar{p}_{it}^g$ where $a_p \in [0, 1]$ is a variable in the action space. It regulates the amount of active power curtailment.

Under the strategy with rectangular action space shown in Fig. 3(b), the reactive power injected/absorbed by inverter i is limited by the active power capacity of the inverter. It can be expressed by $|q_{it}^g| \le \bar{q}_{it}^{gR}$ where $\bar{q}_{it}^{gR} = \sqrt{\bar{S}_i^2 - (\bar{p}_{it}^g)^2}$. We rewrite the equation as $q_{it}^g = a_q \bar{q}_{it}^{gR}$, where $a_q \in [-1, 1]$ is another variable in the action space. It controls the reactive power set point of the inverter.

Under the control strategy with polar action space shown in Fig. 3 (c), the reactive power injected/absorbed by inverter i is limited by the active power provided by inverter. It can be expressed by $|q_{it}^g| \leq \bar{q}_{it}^{gP}$ where $\bar{q}_{it}^{gP} = \sqrt{\bar{S}_i^2 - (p_{it}^g)^2}$. We rewrite the equation as $q_{it}^g = a_q \bar{q}_{it}^{gP}$ where $a_q \in [-1, 1]$.

c) Reward: The reward received by the reinforcement learning agent consists of four terms as shown in (16): line loss, voltage violations, active power curtailment cost, and the inverter degradation cost. The line losses, voltage deviation losses, and the active power curtailment cost are formulated in the same way as in Section III-C. The inverters include power switching devices such as insulated gate bipolar transistors (IGBTs) and diodes. Change in the real and reactive power injection by the smart inverters leads to temperature swings in the switching components which can cause additional thermal stresses, ultimately leading to a reduction of the inverter lifetime. Therefore, we model the inverter degradation cost proportional to the change in the real and reactive power levels of the inverter in consecutive time steps. Let C_I be the inverter degradation cost ($\$/W$ change in inverter real power and $\$/\text{VAR}$ change in inverter reactive power) and $\mathcal{N}_r$ be the nodes with inverters, then the inverter degradation cost is expressed by $d_t = \sum_{i \in \mathcal{N}_r} C_I \left(\left| p_{i(t+1)}^g - p_{it}^g \right| + \left| q_{i(t+1)}^g - q_{it}^g \right| \right)$.

The reward at time t then can be written as follows:

$$r_t = - \sum_{(i,j) \in \mathcal{L}} C_e r_{ij} \ell_{ijt} - \sum_{i \in \mathcal{N}} C_v \left(u_{it} - 1 \right)^2 - \sum_{i \in \mathcal{N}_r} C_c \left| \bar{p}_{it}^g - p_{it}^g \right| - d_t \quad (16)$$

B. Deep Deterministic Policy Gradient

The fast timescale VVC by smart inverters has a continuous and high dimensional action space. In addition, the complete distribution feeder parameters are not always available. Thus, we adopt the deep deterministic policy gradient (DDPG) algorithm [40], a model-free approach, to solve the fast timescale VVC problem. DDPG is an off-policy deep reinforcement learning algorithm with the actor-critic architecture and function approximators. As such, both policy and value functions are approximated by deep neural networks. The actor-network maintains a deterministic policy μ using a neural network parameterized by θ^μ. The input of the neural network is the state s and the output is a deterministic continuous action $a = \mu(s|\theta^\mu)$. To ensure exploration, noise sampled from a noise process η, e.g., an Ornstein-Uhlenbeck process [41] is added to the output: $\mu'(s_t) = \mu(s_t|\theta_t^\mu) + \eta$. The critic network approximates the corresponding Q function of the policy using the neural network parameterized by θ^Q. To improve the stability of learning, two target networks $Q'\left(s, a|\theta^{Q'}\right)$ and $\mu'\left(s|\theta^{\mu'}\right)$ are introduced to provide stable learning targets. As such, the update equations of the network are not interdependent on the values calculated by the network itself and therefore are not prone to divergence.

To further stabilize the training process, the experience replay mechanism is employed to break the correlations between the training experiences: the experience tuples (s_t, a_t, r_t, s_{t+1}) are stored in a replay buffer. Then, random mini-batches of experience are sampled from the replay buffer while updating the value and policy networks.

Since the target policy is deterministic, the Bellman equation can be expressed as follows:

$$Q^\mu(s_t, a_t) = \mathbb{E}\left[R(s_t, a_t) + \gamma \left[Q^\mu(s_{t+1}, \mu(s_{t+1})) \right] \right] \quad (17)$$

The training of the critic network is based on minimizing the following loss function using batches of experiences with N_m number of transitions.

$$L = \frac{1}{N_m} \sum_i \left(y_i - Q(s_i, a_i)|\theta^Q \right)^2 \quad (18)$$

$$y_i = R(s_i, a_i) + \gamma Q'\left(s_{i+1}, \mu'\left(s_{i+1}|\theta^{\mu'} \right) |\theta^{Q'} \right) \quad (19)$$

The parameters of the actor network are updated using the critic network and the policy gradient algorithm with batches of experience with N_m transitions.

$$\nabla_{\theta^\mu} J \approx \frac{1}{N_m} \sum_i \nabla_a Q\left(s, a|\theta^Q\right)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_i} \quad (20)$$

C. Summary of the Two-timescale VVC Algorithm

First, the slow timescale control problem is solved for each hour τ using (1)-(7) to determine the tap positions of the voltage regulators, OLTCs, and switchable capacitor banks. Within the interval τ, the switching decisions of these devices are kept fixed. Now, the fast timescale control of the smart inverters is performed for each time segment t within τ. The corresponding distribution system voltage at each bus along with the load and PV generation and time stamp data is utilized to assemble the state vector s_t for the DDPG training. The state vector is provided to the agent which generates the suggested actions, i.e., the real and reactive power outputs of the inverters. The suggested actions are executed in the environment and the agent gathers the state variables from the environment which transitions to the next time period s_{t+1}. The transition (s_t, a_t, r_t, s_{t+1}) is stored. The actor and critic network is updated following Section IV-B utilizing target network and experience replay till the terminal state is reached. After completing the training, the trained DDPG agent can be utilized to determine the real and reactive power setpoints of the smart inverters for the fast timescale VVC. The detailed process for the two-timescale VVC is shown in Algorithm 1.

V. Numerical Study

The performance of the proposed two-timescale VVC in Algorithm 1 is tested on a modified IEEE 34 node test feeder.

A. Simulation Setup

As shown in Fig. 1, the IEEE 34-bus test feeder has a voltage regulator at node 800. There are two OLTCs connecting node 814 to node 850 and node 852 to node 832 respectively. Two capacitors are placed at node 844 (100 kVar) and node 847 (150 kVar). Three solar PV systems with nameplate

Algorithm 1 Two-timescale Volt-VAR Control scheme

Input: Initial actor network parameters θ^{μ}, critic network parameters θ^{Q}, and empty replay buffer D. Initialize a random process η for action exploration

Initialize target network parameters: $\theta^{\mu'} \leftarrow \theta^{\mu}, \theta^{Q'} \leftarrow \theta^{Q}$

for $t = 1 \dots T \times \mathcal{T}$ **do**

  Fix the tap positions of the conventional voltage regulating devices at the solution obtained from (1)-(7) at the corresponding hour τ

  Obtain load, solar PV generation and voltage magnitude information at time t to form state vector s_t.

  Feed the state vector into the actor network to generate suggested actions, i.e., the real and reactive power outputs of inverters. Select action $a_t = \mu\left(s_t|\theta^{\mu}\right) + \eta_t$ according to current policy and exploration noise.

  Execute a_t in the environment.

  Gather information for the next state s_{t+1}. Calculate the reward r_t.

  Store (s_t, a_t, r_t, s_{t+1}) in replay buffer $\mathcal{D}$

  Randomly sample a batch of N_m transitions from $\mathcal{D}$, $B = \{(s_i, a_i, r_i, s_{i+1})\}$

  Compute target y_i using (19)

  Update Q-function by minimizing loss in (18)

  Update policy by one step of gradient ascent using (20)

  Update target networks with $\theta^{Q'} = \rho\theta^{Q} + (1-\rho)\theta^{Q'}$ and $\theta^{\mu'} = \rho\theta^{\mu} + (1-\rho)\theta^{\mu'}$

end for

capacity 22 KW, 67 KW, and 133 KW are added to the feeder at the nodes 840, 862, and 838 respectively. The inverters are not oversized. The solar PV penetration level of the feeder is 120%. To illustrate the algorithm's capability for active power curtailment and reactive power absorption under low load and high PV production conditions, we double the line impedances so that the benefits of active power curtailment and reactive power absorption are more pronounced.

All voltage regulators and on-load tap changers have 11 tap positions, which correspond to turns ratios ranging from 0.95 to 1.05. The capacitors can be switched on/off remotely and the number of 'tap positions' is treated as 2. In the initial state, the turns ratios of voltage regulators and on-load tap changers are 1 and the capacitors are switched off. The electricity price C_e is assumed to be $\$40/MWh$. The operating cost per tap change is set to be $\$0.1$ for all devices. The penalty coefficient C_V is set as $\$1/volt$. The inverter degradation cost C_I is set to be $\$0.04/MW$. One year of hourly smart meter energy consumption data from London [42] is used. The aggregated load data is scaled and allocated to each node according to the existing spatial load distribution of the IEEE 34-bus test feeder. One year of solar PV generation data from Austin, Texas in 2019 is obtained from the Pecan Street Dataset [43] and scaled according to the corresponding nameplate capacity of the solar PV systems. Five weeks of data from the 9th to 13th week is used for training, in which the agent interacts with the environment and updates its policy and value networks. One week of data for week 14 is used for out-of-sample testing, in which the trained reinforcement learning agent takes control actions without further updating its neural networks.

TABLE I: Hyperparameter settings for DDPG

Parameters	Value
Size of hidden layers	$(512, 512)$
Activation function	ReLU
Batch size	100
Discount factor	0.99
Learning rate actor and critic network	0.0001
Epoch	2
Start steps before running policy	100
Standard deviation for exploration noise	0.4

B. Setup of the Benchmark and Our Proposed Algorithms

Under the model-free reinforcement learning-based control framework, we compare our proposed DDPG-based smart inverter control with polar action space with two other benchmark reinforcement learning algorithms, which have a reactive power control strategy and a real and reactive power control strategy with a rectangular operating space, respectively. In addition, we consider three baseline control scenarios under the model-based control assuming the accurate and complete distribution network model is available.

The three baseline control scenarios under the model-based control framework are set up as follows:

1) Baseline 1: No Volt-VAR control is executed.
2) Baseline 2: Only slow timescale VVC is executed following the method in Section III-B. The smart inverters operate at unity power factor with no reactive power injection/absorption or active power curtailment.
3) Baseline 3: Slow timescale VVC is executed following the method in Section III-B. The smart inverters are controlled following the method in Section III-C.

The slow timescale VVC is formulated as a mixed-integer nonlinear programming (MINLP) and solved by the BONMIN solver in the OPTI toolbox [44] in MATLAB. The optimization-based fast timescale inverter control in baseline scenario 3 is implemented using the CVX toolbox [45] in MATLAB after the convex relaxation is performed.

The setup of our proposed two-timescale Volt-VAR control scheme with three different action space are discussed below:

1) DDPG with only reactive power control: Slow timescale VVC is executed following Section III-B. The smart inverters are controlled using DDPG with only adjustable reactive power setpoint as depicted in Fig. 3(a).
2) DDPG with rectangular action space: Slow timescale VVC is executed following Section III-B. The smart inverters are controlled using DDPG with rectangular action space for real and reactive power setpoints as depicted in Fig. 3(b).

TABLE II: Comparison of the operation costs of the proposed two-timescale VVC schemes along with three baseline scenarios in the test dataset

Operational cost ($)	Baseline 1 (no VVC)	Baseline 2 (slow time scale VVC)	Baseline 3 (optimization based two timescale VVC)	DDPG and reactive power based two timescale VVC	DDPG based two timescale VVC with rectangular action space	DDPG based two timescale VVC with polar action space
Switching	0.00	38.20	38.20	38.20	38.20	38.20
Line loss	33.78	81.30	127.47	169.40	150.60	115.79
Voltage deviation	3264.66	1118.60	352.73	436.12	410.18	414.47
APC	0.00	0.00	14.38	0.00	13.05	16.95
Inverter degradation	0.75	0.75	4.16	4.10	2.80	2.37
Total	3299.20	1238.86	536.96	648.44	614.85	587.78

3) DDPG with polar action space: Slow timescale VVC is executed following Section III-B. The inverters are controlled using DDPG with polar action space for real and reactive power setpoints as depicted in Fig. 3(c).

The feedforward neural networks of both actor and critic networks have 2 fully connected hidden layers of 512 neurons each. At the start of the training, uniform-random actions are selected before running the real policy to help exploration. The training of the agent is performed for 2 epochs. A n epoch refers to one cycle through the full training dataset. The hyperparameter settings for the DDPG algorithm of all three control strategies are provided in Table I.

C. Result and Analysis

To evaluate the performance of the proposed reinforcement learning-based VVC methods, we compute the line loss, voltage violation cost, active power curtailment cost (APC), switching cost of the conventional voltage regulating devices, inverter degradation cost, and the total operational cost. A lower total operational cost indicates a better control performance in voltage regulation. Table II shows the operational cost comparison of three variations of the proposed reinforcement learning-based two-timescale VVC algorithm with three model-based baseline control scenarios on the test dataset. The result is based on the trained model, which achieves the best performance out of 20 random experiments in the training dataset.

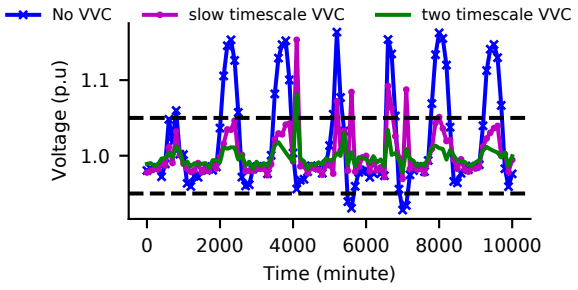

Fig. 4: Comparison of voltage deviations at node 838 for three VVC schemes

It can be observed from Table II that although the slow timescale VVC (Baseline 2) provides voltage regulation service, it is not adequate as the rapid change in the solar PV production within each hour causes high voltage violation cost. As shown in column 4-7 of Table II, all of the two-timescale VVC schemes achieved considerably lower total operational cost. In particular, among the DDPG-based smart inverter control schemes, the proposed two-timescale VVC with polar action space yields the lowest operational cost.

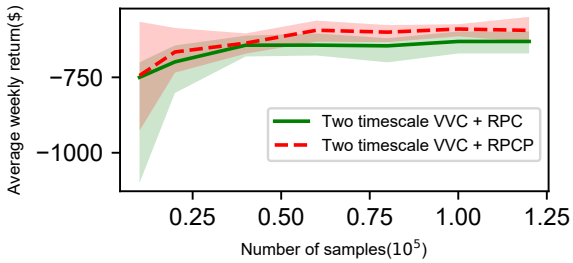

Fig. 5: AVR vs number of samples for the two-timescale VVC schemes with polar action space and with only adjustable reactive power

The DDPG-based two-timescale VVC with polar action space has a larger reactive power adjustment range than that of the control strategy with rectangular action space as shown in Fig 3. Thus, the reinforcement learning-based control with polar action space provides better voltage regulation service and consequently lower operation costs. Although the model-based fast timescale inverter control together with the slow timescale control offers the lowest operation cost, it requires complete and accurate knowledge of the secondary distribution circuit model and parameters. The DDPG-based fast timescale control on the other hand is model-free and produces relatively low total operational cost. The cost for active power curtailment and inverter degradation during the six week training period is $633 for the customers of the 34-bus distribution network as opposed to $91.71 if the optimization from Baseline 3 is implemented. Such training costs are inevitable in reinforcement learning based VVC methods involving smart inverter control as there are exploratory actions.

Next, we compare the voltage profiles of two baseline control scenarios and our proposed DDPG-based VVC with polar action space. The voltage magnitude time series of node 838 corresponding to no VVC, only slow timescale VVC, and the proposed two-timescale VVC with polar action space are shown in Fig. 4. Node 838 is selected

for the comparison because it experiences the worst voltage violation when no VVC is employed. It can be seen that our proposed DDPG-based VVC with polar action space significantly improves the voltage regulation performance. Furthermore, our proposed two-timescale DDPG-based VVC is capable of maintaining the voltage within 1 ± 0.05 p.u. for almost the entire operating week.

Finally, the RL algorithm employed to solve the VVC problem should be sample efficient and scalable. We demonstrate the sample efficiency of the proposed DDPG-based two-timescale VVC algorithm. The number of training samples collected versus the average weekly return (AVR) on the testing weeks are shown in Fig. 5. The AVR is defined as the summation of all the components of the reward function accumulated over the testing period. The middle curve shows the mean AVR averaged over 10 independent runs. The light-colored region corresponds to the error bounds. Fig. 5 also demonstrates the sensitivity of the test set results to the training sequence. It is observed that with about three weeks of training data, the algorithm is able to learn a very effective VVC policy. It should be noted that in Fig. 5, each point on the horizontal axis corresponds to a "training set", which consists of the data from the beginning up to that point, whereas the testing dataset always starts from week 14. Thus the latter does not immediately follow the end of the training dataset. This further shows the effectiveness of the algorithm on out-of-sample data. In addition, as shown by the error bound, these results are consistent across different random initialization and training sessions.

VI. Conclusion

A two-timescale Volt-VAR control scheme that does not depend on accurate secondary feeder models is proposed in this paper. In the slow timescale control, tap positions of conventional voltage regulating devices, such as the voltage regulator, on load tap changers, and switchable capacitor banks are determined by a model-based controller. On the fast timescale, a DDPG-based algorithm is developed to determine the real and reactive power setpoints of the smart inverters. The proposed algorithm is relatively safe to implement in the real world as the slow timescale VVC devices are set according to an optimization based approach; only the smart inverters are allowed to perform exploratory actions. As shown in the numerical study, there is no severe voltage violation during the training period. The proposed DDPG-based smart inverter control strategy with polar action space outperforms the strategy with the rectangular action space and the strategy with only adjustable reactive power. It is capable of maintaining the voltage within a reasonable range. In addition, it is very sample efficient and only requires three weeks of training data to achieve near-optimal results. In the future, we plan to make the entire two-timescale VVC framework model-free. This way, the VVC does not even depend on an accurate primary feeder model in power distribution systems, which may not always be available.

References

[1] EIA, "Annual energy outlook 2020: with projections to 2050," *United States Energy Information Administration*, 2020.

[2] W. Wang, N. Yu, and R. Johnson, "A model for commercial adoption of photovoltaic systems in California," *Journal of Renewable and Sustainable Energy*, vol. 9, no. 2, p. 025904, 2017.

[3] G. Wang, V. Kekatos, A. J. Conejo, and G. B. Giannakis, "Ergodic energy management leveraging resource variability in distribution grids," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4765–4775, Nov. 2016.

[4] B. Mather, S. Shah, B. Norris, J. Dise, L. Yu, D. Paradis, F. Katiraei, R. Seguin, D. Costyk, J. Woyak, J. Jung, K. Russell, and R. Broadwater, "NREL/SCE high penetration pv integration project: FY13 annual report," National Renewable Energy Laboratory (NREL), Golden, CO (United States), Tech. Rep. NREL/TP-5D00-61269, 1136232, Jun. 2014. [Online]. Available: http://www.osti.gov/servlets/purl/1136232/

[5] "IEEE Standard for interconnection and interoperability of distributed energy resources with associated electric power systems interfaces–Amendment 1: To provide more flexibility for adoption of abnormal operating performance category III," *IEEE Std 1547a-2020 (Amendment to IEEE Std 1547-2018)*, pp. 1–16, Apr. 2020.

[6] M. Farivar, C. R. Clarke, S. H. Low, and K. M. Chandy, "Inverter VAR control for distribution systems with renewables," in *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. Brussels, Belgium: IEEE, Oct. 2011, pp. 457–462. [Online]. Available: http://ieeexplore.ieee.org/document/6102366/

[7] H.-G. Yeh, D. F. Gayme, and S. H. Low, "Adaptive VAR control for distribution circuits with photovoltaic generators," *IEEE Transactions on Power Systems*, vol. 27, no. 3, pp. 1656–1663, Aug. 2012.

[8] H. Xu, A. D. Domínguez-García, and P. W. Sauer, "Optimal tap setting of voltage regulation transformers using batch reinforcement learning," *IEEE Transactions on Power Systems*, vol. 35, no. 3, pp. 1990–2001, 2019.

[9] K. Turitsyn, P. Šulc, S. Backhaus, and M. Chertkov, "Distributed control of reactive power flow in a radial distribution circuit with high photovoltaic penetration," *IEEE PES general meeting*, pp. 1–6, 2010.

[10] K. Turitsyn, P. Sulc, S. Backhaus, and M. Chertkov, "Options for control of reactive power by distributed photovoltaic generators," *Proceedings of the IEEE*, vol. 99, no. 6, Jun. 2011.

[11] E. Dall'Anese, S. V. Dhople, B. B. Johnson, and G. B. Giannakis, "Decentralized optimal dispatch of photovoltaic inverters in residential distribution systems," *IEEE Transactions on Energy Conversion*, vol. 29, no. 4, pp. 957–967, 2014.

[12] D. K. Molzahn, F. Dörfler, H. Sandberg, S. H. Low, S. Chakrabarti, R. Baldick, and J. Lavaei, "A survey of distributed optimization and control algorithms for electric power systems," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2941–2962, Nov. 2017.

[13] K. E. Antoniadou-Plytaria, I. N. Kouveliotis-Lysikatos, P. S. Georgilakis, and N. D. Hatziargyriou, "Distributed and decentralized voltage control of smart distribution networks: Models, methods, and future research," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2999–3008, Nov. 2017.

[14] K. Turitsyn, P. Sulc, S. Backhaus, and M. Chertkov, "Local control of reactive power by distributed photovoltaic generators," in *2010 First IEEE International Conference on Smart Grid Communications*, Oct. 2010, pp. 79–84.

[15] P. Jahangiri and D. C. Aliprantis, "Distributed Volt/Var control by PV inverters," *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 3429–3439, Aug. 2013.

[16] M. Farivar, L. Chen, and S. Low, "Equilibrium and dynamics of local voltage control in distribution systems," in *52nd IEEE Conference on Decision and Control*, Dec. 2013, pp. 4329–4334.

[17] R. Tonkoski, L. A. C. Lopes, and T. H. M. El-Fouly, "Coordinated active power curtailment of grid connected PV inverters for overvoltage prevention," *IEEE Transactions on Sustainable Energy*, vol. 2, no. 2, pp. 139–147, Apr. 2011.

[18] R. Tonkoski and L. A. C. Lopes, "Impact of active power curtailment on overvoltage prevention and energy production of PV inverters connected to low voltage residential feeders," *Renewable Energy*, vol. 36, no. 12, pp. 3566–3574, 2011.

[19] E. Dall'Anese, S. V. Dhople, and G. B. Giannakis, "Optimal dispatch of photovoltaic inverters in residential distribution systems," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 2, pp. 487–497, Apr. 2014.

[20] B. A. Robbins, C. N. Hadjicostis, and A. D. Domínguez-García, "A two-stage distributed architecture for voltage control in power distribution systems," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1470–1482, May 2013.

[21] 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*, Jul. 2012, pp. 1–7.

[22] Y. Xu, Z. Y. Dong, R. Zhang, and D. J. Hill, "Multi-timescale coordinated Voltage/Var control of high renewable-penetrated distribution systems," *IEEE Transactions on Power Systems*, vol. 32, no. 6, pp. 4398–4408, Nov. 2017.

[23] C. Li, V. R. Disfani, H. V. Haghi, and J. Kleissl, "Optimal voltage regulation of unbalanced distribution networks with coordination of OLTC and PV generation," in *2019 IEEE Power Energy Society General Meeting (PESGM)*, Aug. 2019, pp. 1–5.

[24] Y. Gao and N. Yu, "Deep reinforcement learning in power distribution systems: Overview, challenges, and opportunities," in *2021 IEEE power & energy society innovative smart grid technologies conference (ISGT)*. IEEE, 2021, pp. 1–5.

[25] W. Wang, N. Yu, B. Foggo, J. Davis, and J. Li, "Phase identification in electric power distribution systems by clustering of smart meter data," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2016, pp. 259–265.

[26] B. Foggo and N. Yu, "A comprehensive evaluation of supervised machine learning for the phase identification problem," *World Acad. Sci. Eng. Technol. Int. J. Comput. Syst. Eng*, vol. 12, no. 6, 2018.

[27] W. Wang and N. Yu, "Parameter estimation in three-phase power distribution networks using smart meter data," in *2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. IEEE, 2020, pp. 1–6.

[28] B. Foggo and N. Yu, "Improving supervised phase identification through the theory of information losses," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2337–2346, 2019.

[29] 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 Transactions on Smart Grid*, vol. 11, no. 4, pp. 3008–3018, 2020.

[30] Y. Xu, W. Zhang, W. Liu, and F. Ferrese, "Multiagent-based reinforcement learning for optimal reactive power dispatch," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1742–1751, Nov. 2012.

[31] J. G. Vlachogiannis and N. D. Hatziargyriou, "Reinforcement learning for reactive power control," *IEEE Transactions on Power Systems*, vol. 19, no. 3, pp. 1317–1325, Aug. 2004.

[32] C. Li, C. Jin, and R. Sharma, "Coordination of PV smart inverters using deep reinforcement learning for grid voltage regulation," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, Dec. 2019, pp. 1930–1937.

[33] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-timescale voltage regulation in distribution grids using deep reinforcement learning," in *2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, Oct. 2019, pp. 1–6.

[34] H. Xu, A. D. Domínguez-García, and P. W. Sauer, "Data-driven coordination of distributed energy resources for active power provision," *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 3047–3058, 2019.

[35] L. Gan, N. Li, U. Topcu, and S. H. Low, "Exact convex relaxation of optimal power flow in radial networks," *IEEE Transactions on Automatic Control*, vol. 60, no. 1, pp. 72–87, Jan. 2015.

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

[37] S. Huang, Q. Wu, J. Wang, and H. Zhao, "A Sufficient Condition on Convex Relaxation of AC Optimal Power Flow in Distribution Networks," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 1359–1368, Mar. 2017.

[38] Q. Li and V. Vittal, "Non-iterative enhanced SDP relaxations for optimal scheduling of distributed energy storage in distribution systems," *IEEE Transactions on Power Systems*, vol. 32, no. 3, pp. 1721–1732, May 2017.

[39] N. Nazir and M. Almassalkhi, "Voltage positioning using co-optimization of controllable grid assets in radial networks," *IEEE Transactions on Power Systems*, pp. 1–1, 2020.

[40] 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.

[41] G. E. Uhlenbeck and L. S. Ornstein, "On the theory of the Brownian motion," *Phys. Rev.*, vol. 36, pp. 823–841, Sept. 1930.

[42] UK Power Networks, "Smart meter energy consumption data in London households." [Online]. Available: https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households

[43] "Pecan street Inc. Dataport." [Online]. Available: http://www.pecanstreet.org/dataport/

[44] J. Currie and D. I. Wilson, "OPTI: Lowering the Barrier Between Open Source Optimizers and the Industrial MATLAB User," in *Foundations of Computer-Aided Process Operations*, N. Sahinidis and J. Pinto, Eds., Savannah, Georgia, USA, 8–11 January 2012.

[45] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.