Title

Online PV Smart Inverter Coordination using Deep Deterministic Policy Gradient

Permalink

https://escholarship.org/uc/item/1v62w7mw

Authors

Li, Changfu
Chen, Yi-An
Jin, Chenrui
et al.

Publication Date

2022-08-01

DOI

10.1016/j.epsr.2022.107988

Copyright Information

Online PV Smart Inverter Coordination using Deep Deterministic Policy Gradient

Changfu Li, Yi-An Chen, Chenrui Jin, Ratnesh Sharma, and Jan Kleissl

***Abstract*—Fast and frequent solar power variations present new challenges to modern power grid operation with increasing adoption of photovoltaic (PV) energy. PV smart inverters (SIs) provide a fast-response method to regulate voltage by modulating active and/or reactive power at the connection point. In this paper, a deep reinforcement learning (DRL) based algorithm is proposed to coordinate multiple SIs. A reward scheme is designed to balance voltage regulation and SI reactive power utilization. The proposed DRL agent for voltage control learns its policy through massive offline simulations and adapts to load and solar variations. The DRL agent results are compared against autonomous Volt-Var control and optimal power flow (OPF) on the IEEE 37 bus feeder and IEEE 123 bus feeder for 8760 different scenarios. The results demonstrate that a properly trained DRL agent can intelligently coordinate different SIs to satisfy grid voltage limits despite large solar and load variations. The DRL agent achieves nearly the optimal performance of OPF by mitigating all voltage violations, while reducing PV production curtailment by 88% compared to the autonomous Volt-Var scheme. Contrary to OPF, the DRL agent can provide a coordination signal in milliseconds without load and solar forecasts and without explicit knowledge of the distribution network model.**

***Keywords*-Deep Reinforcement Learning; Distribution Network; Photovoltaics; Voltage Regulation; Smart Inverter**

NOMENCLATURE

ΔI_k	current mismatch at node k
γ	future reward discount factor
$\lvert V_k \rvert$	voltage magnitude of node k
$\mathcal{A}$	action space
$\mathcal{S}$	state space
μ	actor function
π	agent policy
θ^{μ}	parameters of actor function
θ^{Q}	parameters of critic function
θ_k	voltage angle of node k
a_t	action at time step t
B_{kj}	susceptance between node k and j
C	positive constant in the term of R_Q
E	environment of reinforcement learning
G_{kj}	conductance between node k and j
I_k^{calc}	calculated current injection at node k
I_k^{sp}	specified current injection at node k
I_k	current phasor of node k
M	dimension of action space
N	total number of non-slack nodes
P_i^{pv}	active power generation of i^{th} PV
P_k^{calc}	calculated active power injection at node k
P_k^{l}	active load power at node k
P_k^{sp}	specified active power injection at node k
P_k	net active power injection at node k
$Q(s,a)$	critic function
q_i	reacive power utilization ratio of i^{th} SI
Q_i^{pv}	reactive power generation of i^{th} PV/SI
Q_k^{calc}	calculated reactive power injection at node k
Q_k^{l}	reactive load power at node k
Q_k^{sp}	specified reactive power injection at node k
Q_k	net reactive power injection at node k
R_Q	reward associated with reactive power usage
R_t	total discounted future reward from time step t and onwards
r_t	scalar reward for time step t
R_V	reward associated with voltage
S_i	power rating of i^{th} SI
s_t	state of the environment at time step t
V_{slack}^{sp}	specified voltage for slack bus nodes
V_k	voltage phasor of node k
V_k^{nom}	nominal voltage of node k
V_{slack}	voltage of slack bus nodes

I. INTRODUCTION

Renewable distributed generation (DG) adoption has seen significant increases recently due to the associated technical, economic, and environmental benefits [2]. DG adoption also presents various new challenges to grid operators. For instance, voltage violations can become a problem due to increasing penetration of variable DGs such as solar photovoltaic (PV) generators [3, 4].

Conventionally, distribution network operators rely on on-load tap changers (OLTCs) and fixed or switched shunt capacitors/reactors to maintain appropriate voltages across the network. OLTCs typically work in autonomous mode, following simple pre-defined rules based on local measurements. This simple voltage regulation scheme is effective for conventional centralized power supply with monotonously decreasing voltage profile along the feeder and slow voltage changes. However, the presence of large amount of renewable DGs on the distribution network can cause reverse power flow, leading to voltage increases along the feeder and possible voltage violations [5]. Due to their electro-mechanical nature, OLTCs

A preliminary version of this paper was published in the 18th IEEE International Conference on Machine Learning and Applications - ICMLA 2019, Boca Raton, FL, 2019 [1].

Changfu Li, Yi-An Chen and Jan Kleissl are with the Center for Energy Research and the Department of Mechanical and Aerospace Engineering at the University of California San Diego, La Jolla, CA 92093 USA. {emails: chl447@ucsd.edu, yic002@ucsd.edu, jkleissl@ucsd.edu}

Portions of the work were done when Chenrui Jin and Ratnesh Sharma were with NEC Labs America, Inc., San Jose, CA USA. {email: jinchenrui@gmail.com, ratnesh_sharma@yahoo.com}

are limited in the number of tap changes before preventive maintenance or overhaul is required. To reduce the number of tap changes, OLTCs are programmed to act with delays of 10s of seconds and therefore respond slowly. Therefore, OLTCs are less effective in controlling voltage with sub-minute PV variability. On the contrary, smart inverters (SIs) can rapidly respond to voltage regulation by modulating active and/or reactive power of PV systems at the point of common coupling (PCC) [6]. The commonly used SI Volt-Var functions, as defined in [7–9], are based on local droop curves, which define the SI absorption/injection of reactive power according to the local bus voltage. Local droop curves result in sub-optimal system performance due to lack of coordination.

Various studies have aimed to improve basic rule-based autonomous local voltage control schemes [10–12]. Reference [10] improve the sensitivity of voltage control to downstream voltage using feeder end measurements instead of local bus voltage for the control of tap switching. Voltage estimates from a sensitivity matrix are adopted to dynamically adjust OLTC voltage set points to accommodate SI outputs in [11]. OLTC and SIs are coordinated by iteratively updating their settings to achieve target voltages at the SIs in [12].

These improved rule-based voltage control methods are relatively simply to implement and can achieve partial coordination between different devices. However, optimization-based approaches can realize optimal voltage regulation to combat complicated voltage profiles caused by renewable DGs. Some previous works focus on coordination of SIs [13–16]. Optimal power flow (OPF) is formulated as a second-order cone program to optimize SI reactive power for line loss reduction while meeting voltage requirements [13]. Reference [14] adopts the alternating direction method of multipliers (ADMM) to solve the OPF problem and find the optimal SI reactive power to reduce losses. ADMM-based algorithms are also employed in [15] to determine SI active and reactive power set points for voltage regulation. Reference [16] applies dynamic weight-based collaborative optimization to dispatch SI reactive power for voltage deviation minimization.

Other works optimize cooperation between other devices as well as SIs [17–22]. Reference [17] proposed a linearization model to optimize multiple OLTCs to minimize voltage deviations. OLTC tap positions and SI outputs (reactive power in [18, 19]; active and reacive power in [20]) are optimized concurrently to minimize voltage deviations. DGs and OLTCs are coordinated through optimization to minimize voltage deviations and network losses in [21]. SIs, OLTCs, and shunt capacitors are coordinated in [22] to meet voltage operation limits.

Despite optimization-based approaches and OPF methods accomplishing optimal voltage regulation, there are two major limitations. First, they require accurate distribution network models including resistances and reactances, and the network topology, which are not necessarily available [23]. Second, non-linear power flow constraints render the optimization problem computationally intensive, especially for large networks. Long run times limit the methods' practical application to address fast PV disturbances caused by moving clouds.

The success of reinforcement learning (RL), especially deep reinforcement learning (DRL) in various fields including AlphaGo [24] and robotics [25], has attracted interest in the power and energy community. Numerous works have applied RL/DRL for intelligent control and operation in power grids. Deep Q network (DQN) and Deep Deterministic Policy Gradient (DDPG) are used in [26, 27] for controlling discrete generator voltage set points to maintain acceptable system voltages in response to load variations and line outages. The double Deep Q network (DDQN) is applied to optimal active power dispatch to achieve operation cost reduction in [28]. DDQN is also adopted in [29] to control grid topology changes to maximize available transfer capabilities. Reference [30] uses a multi-agent DDPG method to adjust generator voltage set points continuously to solve the classic autonomous voltage control problem in the transmission grid. Reference [31] dispatches SIs, OLTCs, and capacitors at two timescales for distribution network voltage control: An optimization is used for fast dispatch of SIs while slow OLTCs and capacitors are handled by DQN. Batch RL is applied to achieve cooperation of OLTCs for voltage regulation [32]. Coordination between OLTCs and capacitors [33–35] and SI reactive power dispatch [36] are studied with the policy gradient method for voltage violation mitigation and operation cost reduction. Multi-agent DRL is used in [37, 38] to dispatch SI reactive power and static Var compensators to address voltage violations. Reference [39] coordinates OLTCs, capacitors, and generators to meet operation limits with Q-learning.

Of the works that use RL for power grid applications, references [31–39] focus on distribution voltage regulation. Within those, [31–35, 39] use RL/DRL only for control of legacy voltage regulation devices with discrete settings (i.e. generator voltage set points, OLTC tap positions, capacitor switches).

SIs are more suitable for mitigating frequent PV generation fluctuations due to their continuous outputs and fast response in comparison to legacy voltage regulation devices. While references [36–38] coordinate SIs with DRL, PV active power curtailment is not considered in the reward function design in [37, 38], which can lead to excessive curtailment. Reference [36] balances active power curtailment and voltage regulation. However, instead of directly determining optimal active and reactive power set points, incremental changes are employed, which can lead to insufficient responses to large PV ramps. Moreover, the performance is not validated against OPF.

In this paper, we propose a DDPG-based algorithm to coordinate multiple SIs with continuous outputs. The reward is carefully designed to balance voltage regulation and active power curtailment, in contrast to [37, 38]. Unlike OPF approaches, the proposed DDPG agent is data-driven and relies on little to no knowledge of the distribution network. The DDPG approach can reach decisions in milliseconds, fully leveraging the fast-response speed of SIs to deal with frequent and fast solar ramps. The DDPG method is validated against the autonomous Volt-Var scheme [8] and OPF (contrary to [36]) on the modified IEEE 37 bus feeder and the IEEE 123 bus feeder. Comprehensive tests are carried out for a full year (8760 different scenarios) to demonstrate the effectiveness and robustness of the well-trained DDPG agent.

The rest of the paper is organized as follows. Section II introduces preliminaries of the distribution network and SIs. The OPF formulation and DDPG implementation for coordination of SIs is presented in Section III. Case studies are detailed in Section IV. Results and discussion are presented in Section IV-C, followed by conclusions in Section VI.

II. Preliminaries

A. Distribution System

From the graph theory perspective, a distribution network with $N+1$ nodes can be represented by a graph $\mathcal{G} := (\mathcal{N}_0, \xi)$, where $\mathcal{N}_0 := \{0, ..., N\}$ is the collection of all nodes, and $\xi := \{(m, n) \subset \mathcal{N}_0 \times \mathcal{N}_0\}$ is the collection of edges representing distribution lines. The distribution network typically operates radially as a tree and is served by a substation (a.k.a. the root) indexed by $n = 0$. The primary side of substation can be treated as a slack bus, where voltage magnitude $|V_0|$ and angle θ_0 can be modeled as constants. The voltage for all $N+1$ nodes is governed by the power flow equations:

$$\sum_{j=0}^{N} |V_k||V_j|\big(G_{kj}cos(\theta_k - \theta_j) + B_{kj}sin(\theta_k - \theta_j)\big) - P_k = 0 \tag{1}$$

$$\sum_{j=0}^{N} |V_k||V_j|\big(G_{kj}sin(\theta_k - \theta_j) - B_{kj}sin(\theta_k - \theta_j)\big) - Q_k = 0 \tag{2}$$

where $|V_k|$ and θ_k are the voltage magnitude and voltage angle at node k, respectively; G_{kj} and B_{kj} are the conductance and susceptance of the electrical line connecting nodes k and j; P_k and Q_k are the net active and reactive power injections at node k.

B. Smart Inverter for Voltage Regulation

A PV inverter is a type of electrical device that converts the direct current (DC) output of solar panels into alternating current (AC), which can then be fed into the AC grid through the point of common coupling (PCC). Under the new standards/rules [7–9], PV inverters are required to contribute to grid regulation via defined functions; this type of PV inverter is referred to as a smart inverter (SI). A SI supports voltage regulation by modulating active and/or reactive power at the PCC; in other words, the SI at node k can change the P_k and/or Q_k in (1,2) affecting the voltage for node k as well as other nodes, per (1,2).

A commonly used smart function is the Volt-Var droop curve, as shown in Fig. 1. Six unique points specify the shape of the curve, according to which the SI will absorb or inject the corresponding amount of reactive power (var) based on the voltage at the PCC. The PV active power production can be curtailed to make headroom for var generation if the SI reaches its capacity limit as shown in Fig. 2. This scheme is called Volt-Var with var priority. With a Volt-Var droop curve, every SI operates autonomously (i.e. without coordination with other inverters) based on its local PCC voltage only. While this simplifies the implementation, it can also lead to undesired system performance. For example, since not all nodes of the power network are equipped with SIs, some nodes may suffer from voltage violations. Some SIs may use excessive reactive power due to a lack of coordination with other SIs, resulting in unnecessary PV production curtailment.

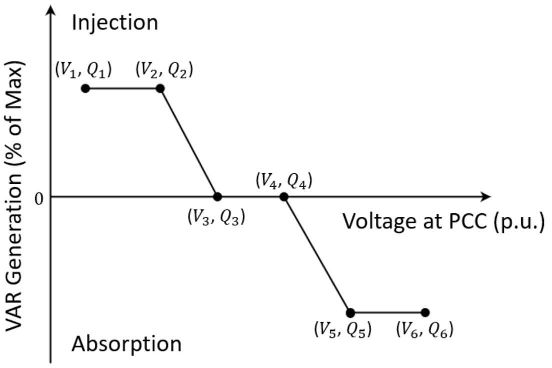

Fig. 1: A typical Volt-Var droop curve of a smart inverter.

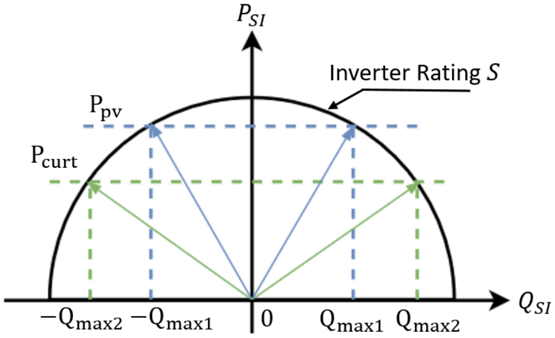

Fig. 2: Smart inverter (SI) output: the complex power output of the SI is $S_{\text{SI}} = P_{\text{SI}} + iQ_{\text{SI}}$. S_{SI} is constrained by the inverter rating S, meaning $P_{\text{SI}}^2 + Q_{\text{SI}}^2 \leq S^2$. P_{pv} is the available PV active power production determined mostly by solar irradiance, $\pm Q_{\max 1}$ is the corresponding maximum reactive power injection or absorption. If the active power is curtailed to P_{curt}, more headroom is created for modulating reactive power ($\pm Q_{\max 2}$).

III. Problem Formulation

A. Reinforcement Learning Formulation

1) Reinforcement Learning: RL, especially DRL, has been shown to be capable of learning by interacting with complicated environments and achieve good performances on difficult control tasks, such as robot manipulation.

The overall RL idea is presented in Fig. 3. An agent learns through interacting with an environment, E. At each time step, the agent receives the state of the environment s_t, takes an action a_t, and receives a scalar reward r_t. The agent learns a policy π, which maps states to a probability distribution over the actions $\pi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$. This can be modeled as a Markov decision process with a state space $\mathcal{S}$, action space $\mathcal{A} = \mathbb{R}^M$, an initial state distribution $p(s_1)$, transition probability $p(s_{t+1}|s_t, a_t)$, and reward function $r(s_t, a_t)$. M is the dimension of the action space.

The agent uses the policy to explore the environment and generate states, rewards, and actions tuples, $(s_1, a_1, r_1, ..., s_t, a_t, r_t)$. The return of a state is calculated as the total discounted future reward from time step t and onwards, $R_t = \sum_{i=t}^{T} \gamma^{(i-t)} r(s_i, a_i)$, where $\gamma \in [0, 1]$ is the discount factor quantifying the importance attached to future rewards. The goal of the agent is to learn a policy that results in maximization of cumulative discounted reward from the start distribution $J = \mathbb{E}_{r_i, s_i \sim E, a_i \sim \pi}[R_1]$.

The action value function is defined as the expected total discounted reward after taking an action a_t in state s_t and thereafter following policy π:

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{r_{i\geq t}, s_{i\geq t} \sim E, a_{i\geq} t \sim \pi}[R_t | s_t, a_t] \tag{3}$$

If the target policy is deterministic, it can be described as a function $\mu : \mathcal{S} \to \mathcal{A}$. The Bellman equation in Q-learning [40] can be expressed as:

$$Q^{\mu}(s_t, a_t) = \mathbb{E}_{r_t, s_{t+1} \sim E}[r_t(s_t, a_t) + \gamma Q^{\mu}(s_{t+1}, \mu(s_{t+1}))] \tag{4}$$

Parameterizing the function approximators with θ^Q, the weights can be optimized by minimizing the loss:

$$L(\theta^Q) = \mathbb{E}\big[(y_t - Q(s_t, a_t | \theta^Q))^2\big] \tag{5}$$

where $y_t = r(s_t, a_t) + \gamma Q(s_{t+1}, \mu(s_{t+1}) | \theta^Q)$.

2) Deep Deterministic Policy Gradient Algorithm: Applying Q-learning (Eq. (4)) to a continuous action space is problematic, as the greedy policy requires global optimization during policy improvement. The deterministic policy gradient (DPG) is more computationally tractable for problems over a continuous action space [41]. The DPG parameterizes the actor with a function $\mu(s|\theta^{\mu})$. The critic $Q(s, a)$ is learned based on the Bellman equation as in Q-learning. Fig. 4 shows the structure of the deterministic actor critic network. The actor is updated via gradient descent to maximize the expected return from the start distribution J:

$$\nabla_{\theta^{\mu}} \approx \mathbb{E}[\nabla_{\theta^{\mu}} Q(s, a | \theta^Q) | s = s_t, a = \mu(s_t | \theta^{\mu})] \tag{6}$$

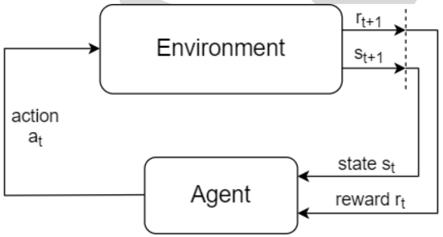

Fig. 3: Schematic overview of reinforcement learning.

In this paper, a similar approach is adapted from [42], which uses deep neural networks as function approximators for DPG. This approach is referred as deep deterministic policy gradient (DDPG).

B. Design DDPG Agent for Smart Inverter Coordination

The goal of a well-trained DDPG agent for SI coordination is to provide fast and effective actions for maintaining voltage limits and minimizing PV production curtailment. The actions are determined based on real-time measurements (states) of the power grid from the supervisory control and data acquisition (SCADA) system or phasor measurement units (PMUs). In this paper, AC power flow (PF) is used to simulate the measurements from SCADA or PMUs. Each node of the feeder is assumed to be equipped with one measurement unit and the PF solution of all nodes are included as states. In actual field applications, to reduce costs, a subset of nodes can be chosen for measurement unit deployment.

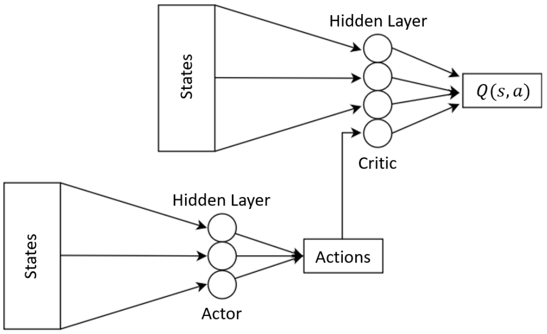

Fig. 4: Deep deterministic policy gradient network. The actor suggests actions based on states. The critic evaluates the actions provided the current states.

The key concepts of episode, states, actions, and rewards are defined below:

1): Episode

The episode is a sequence of interactions between the agent and the environment in response to a specific grid condition/scenario (a combination of PV and load profiles). During an episode, the agent explores by suggesting actions and receiving resultant states and rewards until termination due to convergence or reaching the maximum number of steps. Each exploration step is also referred to as iteration.

2): State Space

The state s is defined as a vector containing power system information, including voltage magnitudes of each node as well as active and reactive power generation/consumption of PVs and loads. The state s belongs to the state space $\mathcal{S}$.

3): Action Space

SI reactive power outputs are actions. Each SI can adjust its reactive power output continuously from $-S$ to S (Fig. 2) or [-1,1] p.u. after normalization, although that may require PV active power curtailment. The action space $\mathcal{A}$ is spanned by action combinations of all SIs.

4): Reward

When applying RL to control, the reward scheme needs to be carefully designed to achieve proper system performance. Since the objectives are mitigating voltage violations and minimizing PV generation curtailment, the reward scheme is composed of two parts: (i) A large penalty for violating voltage limits; and (ii) a negative reward proportional to total reactive power dispatched by SIs. The reward associated with total reactive power is used to achieve regulation for both day time (possible curtailment) and night time (no curtailment).

The first part of the reward is assigned according to all node voltages. Several voltage operation zones are defined (Fig. 5): a normal zone (0.95 - 1.05) p.u., a violation zone 1 (0.9 - 0.95 or 1.05 - 1.1) p.u., and a violation zone 2 ($<$ 0.9 or $>$ 1.1) p.u.. These zones are defined according to the grid operation limits in ANSI standards [43]. Assuming $|V_k|_{norm} = \frac{|V_k|}{V_k^{nom}}$ is the normalized voltage magnitude at node k, where $|V_k|$ is the voltage magnitude at node k and V_k^{nom} is the nominal voltage of node k. The voltage reward associated with $|V_k|_{norm}$ for node k in the j^{th} iteration is:

$$R_V(j,k) = \begin{cases} 0, & \text{if } |V_k|_{norm} \in \text{normal zone} \\ Penalty_1 & \text{if } |V_k|_{norm} \in \text{violation zone 1} \\ Penalty_2, & \text{if } |V_k|_{norm} \in \text{violation zone 2} \end{cases} \tag{7}$$

In other words, the corresponding voltage reward is zero if the node voltage is in the normal zone, and large penalties (i.e. negative rewards) will be assigned if the node voltage is out of the operation limits. In this paper, $Penalty_1$ and $Penalty_2$ are set to be -400 and -600 for the IEEE 37 bus case, while -450 and -650 are used for the IEEE 123 bus case. The values are chosen empirically. Larger penalties will generally result in better voltage regulation performance at a cost of more reactive power utilization and therefore possible higher PV active power curtailment.

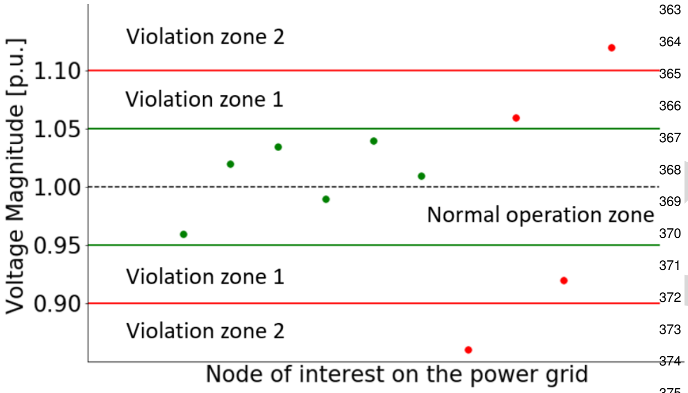

Fig. 5: Definition of voltage profile zones. Each dot represents the voltage of one node. The corresponding total voltage reward for this example is $2 \times Penalty_1 + 2 \times Penalty_2$.

The second part of the reward is assigned based on reactive power utilization. The objective is to minimize PV production curtailment, which is achieved by minimizing reactive power utilization. The reward for reactive power utilization is defined as:

$$R_Q(j) = \sum_{i=1}^{M} C \times (1 - q_i) \tag{8}$$

where $q_i = |Q_i^{pv}|/S_i$ is the reactive power utilization ratio of the i^{th} SI (i.e. the absolute value of action for the SI); M is the total number of SIs/dimension of action space; C is a positive constant chosen to scale the reward.The value of C should be tuned to fit different power system configurations for balancing voltage regulation and reactive power utilization. In this paper, C is empirically set to be 200 and 300 for the IEEE 37 bus and IEEE 123 bus cases, respectively. The total reward for j^{th} iteration/exploration step of the episode is:

$$R(j) = \sum_{k=0}^{N} R_V(j,k) + R_Q(j) \tag{9}$$

where $\{0, ..., N\}$ are the indexes for all nodes.

C. Training of DDPG

DDPG is trained according to the procedures displayed in Fig. 6 with the following key steps:

Step 1: Initialize and solve power flow and assemble state vector: At the beginning of an episode, a new grid operation condition (a combination of PV and load profiles) is randomly generated. The PF is solved to obtain the system information and assemble the state vector. The PF is performed by the AC power flow solver OpenDSS [44], which takes in load consumption and PV generation, solves the corresponding PF equations (1,2), and finds voltages at each node.

Step 2: DDPG agent suggests actions: The state vector containing the system information (node voltage, active and reactive power consumption/generation of SIs and loads) is fed into the DDPG agent. The agent suggests actions, which are SI reactive power outputs.

Step 3: Execute actions and evaluate rewards: The environment (the SI in OpenDSS) take the suggested actions, producing the resultant state by solving another PF. The corresponding reward for that state is evaluated. If the termination criteria defined below are met, the training for this episode is terminated and the trained DDPG is stored for later use. If the termination criteria is not met, we return to **Step 2**, update the agent policy, and repeat the process until the termination is reached.

The training for one episode terminates if: 1) the reward for the exploration/iteration step converges, meaning the reward difference between the current iteration and the last iteration is within 0.5% of the highest reward ($C \times M$ according to (9), i.e. the reward if the agent uses zero reactive power and there are no voltage violations) for five consecutive iterations. To ensure exploration of the search space, at least 200 iterations are performed for each episode independent of whether the convergence criterion is met; or 2) the maximum number of iterations (1,000) is reached.

D. Optimal Power Flow Formulation

To benchmark the performance of the DDPG approach, an equivalent OPF problem is solved. Since the goal is to minimize SI reactive power usage and subsequently PV generation curtailment, the OPF objective is defined as:

$$\min \quad \sum_{i=1}^{M} (-P_i^{\text{pv}} + wQ_i^{\text{pv}}) \tag{10}$$

where M is the total number of PVs/SIs, P_i^{pv} is the PV active power generation, Q_i^{pv} is the reactive power generation, and w is a weighting factor. As shown in Fig. 1, the constraint for every SI is: $(P_i^{\text{pv}})^2 + (Q_i^{\text{pv}})^2 <= S_i^2$, where S_i is the SI power rating. While in theory minimizing either negative active power or reactive power would achieve the goal of minimizing curtailment, in Eq. 10 both terms are included as doing so was found to improved convergence.

The current mismatch equations are used to relate nodal voltages with active and reactive power injections from each load and PV unit. The current mismatch equations are [45]:

$$\Delta I_k = I_k^{\text{calc}} - I_k^{\text{sp}} \tag{11}$$

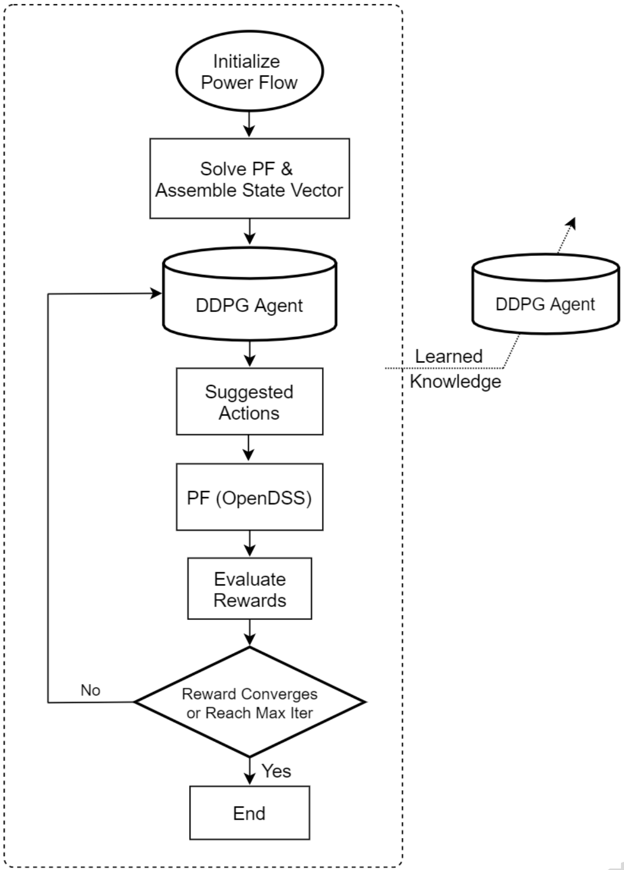

Fig. 6: Flowchart of training a DDPG agent.

$$P_k^{\mathrm{sp}} = \mathrm{Re}(V_k)\,\mathrm{Re}(I_k^{sp}) + \mathrm{Im}(V_k)\,\mathrm{Im}(I_k^{sp}) \tag{12}$$

$$Q_k^{\mathrm{sp}} = \mathrm{Im}(V_k)\,\mathrm{Re}(I_k^{sp}) - \mathrm{Re}(V_k)\,\mathrm{Im}(I_k)^{sp} \tag{13}$$

$$\mathrm{Re}(I_k^{\mathrm{calc}}) = \sum_{j=1}^{N} [G_{ki}\,\mathrm{Re}(V_i) - B_{ki}\,\mathrm{Im}(V_i)] \tag{14}$$

$$\mathrm{Im}(I_k^{\mathrm{calc}}) = \sum_{j=1}^{N} [G_{ki}\,\mathrm{Im}(V_i) + B_{ki}\,\mathrm{Re}(V_i)] \tag{15}$$

where ΔI_k is the mismatch between the calculated current injection I_k^{calc} and the specified current injection I_k^{sp} at node k; V_i and I_i are the voltage phasor (complex) and current phasor for node i; P_k^{sp} and Q_k^{sp} are the specified active and reactive power injections, and G_{ki} and the B_{ki} are the conductance and susceptance from the nodal admittance matrix.

The specified nodal active and reactive power injections at node k are defined by:

$$P_k^{\mathrm{sp}} = P_k^{\mathrm{pv}} - P_k^l \tag{16}$$

$$Q_k^{\mathrm{sp}} = Q_k^{\mathrm{pv}} - Q_k^l \tag{17}$$

where P_k^{pv} is the PV active power injection; P_k^l is the load active power consumption; Q_k^{pv} is the PV reactive power injection, and Q_k^l is the load reactive power consumption.

The OPF formulation forces the current mismatches from (11) to equal zero by imposing the constraint

$$\Delta I_k = 0 \tag{18}$$

The source bus is modeled as a slack bus by constraining its voltage magnitude and angle to be constants determined by the voltage at the primary side of the substation:

$$V_{\mathrm{slack}} = V_{\mathrm{slack}}^{\mathrm{sp}} \tag{19}$$

The voltage constraints at the other nodes are the [0.95, 1.05] p.u. ANSI limits [43]:

$$0.95 \leq \frac{|V_k|}{V_k^{\mathrm{nom}}} \leq 1.05 \tag{20}$$

where V_k^{nom} is the nominal voltage of node k. The OPF solution yields the decision values (P^{pv}, Q^{pv}) to dispatch the SIs.

IV. Case Study

A. Case Study

The proposed DDPG agent is tested on the modified IEEE 37 bus feeder and modified IEEE 123 bus feeder. The properties of the IEEE 37 bus feeder are summarized in Table I. There are 25 loads with a peak load of 2.74 MVA. Five 1.2 MW PVs are randomly deployed with SI AC ratings of 1 MVA, representing a 20% oversizing of the solar array [46]. The resultant PV penetration is around 180%. The IEEE 123 bus feeder (Table II) contains 85 loads with a peak demand of 7.7 MVA. Ten 1.2 MW DC / 1.0 MVA AC PV systems are added at randomly selected locations, achieving 130% PV penetration.

Four different control strategies are tested (Table III): 1) **No Control**: The SI operates at unity power factor without any reactive power generation. 2) **Volt-Var**: Each SI operates autonomously according to the pre-specified local droop curve (Fig. 1) without coordination. 3) **OPF**: All SIs cooperate following the optimal solutions of the OPF problem ((10)). 4) **DDPG**: SIs are coordinated following the decisions made by the trained DDPG agent, as described in Section III.

B. DDPG Agent Training Result

The DDPG agent is trained following Fig. 6. In the training stage, combinations of PV generation and load consumption are randomly generated to present a mix of grid operation conditions/scenarios with low to high loading and PV generation.

Table. IV summarizes neural network topologies used for the IEEE 37 bus and IEEE 123 bus cases. For the IEEE 37 bus case, two hidden layers are used for both actor and critic (Fig. 4) with 200 and 300 neurons. There are 153,210 parameters in total. For the IEEE 123 bus case, a neural network with three hidden layers with 300, 400, and 400 neurons is used. The total number of parameters is 710,020. In general, a larger feeder with more SIs will need larger neural networks (more parameters) to learn a good policy.The neural network topology is customized for the configurations of the feeder. Modifications of neural network topology and retraining are needed if the feeder changes (feeder toplogy, load numbers, PV numbers, etc. Transfer learning [47] techniques can be leveraged to significantly reduce retraining costs.)

Training is performed for 1,500 episodes for the IEEE 37 bus feeder case with 500,000 number of iterations in

TABLE I: Properties of modified IEEE 37 bus feeder.

# of Nodes	37
Peak Load (MVA)	2.74
# of Loads	25
# of PVs	5
Total PV DC Rating (MW)	6
Total PV SI AC Rating (MVA)	5

TABLE II: Properties of modified IEEE 123 Bus feeder.

# of Nodes	128
Peak Load (MVA)	7.7
# of Loads	85
# of PVs	10
Total PV DC Rating (MW)	12
Total PV SI AC Rating (MVA)	10

total (i.e. each episode terminates after approximately 330 iterations, on average). The IEEE 123 bus case is trained for 800 episodes with around 387,000 iterations (i.e. each episode terminates/converges after 484 iterations, on average). For each training episode, a combination of PV and load profiles is randomly generated to represent a grid operation scenario. After training, real PV and load profiles are used to test the trained DDPG agent. No PV and load profile data is shared across training and test stages. Due to the larger and more powerful neural networks (710,020 parameters in total) used for the IEEE 123 bus case, the IEEE 123 bus requires less iterations/data samples to learn a good policy.

The training reward of 5 random experiments for the IEEE 123 bus case is plotted in Fig. 7. The reward R is normalized by the highest possible reward $C \times M$, therefore the reward upper limit here is 1. The reward starts at negative values, given that the grid experiences a large number of voltage violations and the DDPG agent has no prior knowledge on how to perform grid voltage regulation. The DDPG obtains an average reward greater than 0.8 after just 150 episodes, showing that the agent is learning efficiently. As learning progresses further after episode 150, the average rewards remains greater than 0.5 and almost always greater than 0.8 with small swings across different episodes. The episode reward fluctuations are due to differences in grid operation conditions for each episode. Since the grid operation condition/scenario is randomly generated for each episode, some episodes experience grid operation conditions with more violations. Since voltage violations occur in the beginning of these episodes and more reactive power needs to be used to correct voltage violations, lower average episode rewards result.

TABLE III: Summary of smart inverter control strategies.

	SI Reactive Power	SI Dispatch Scheme
No Control	No	N/A
Volt-Var	Yes	local droop curve (Fig. 1)
OPF	Yes	coordinated by OPF solution
DDPG	Yes	coordinated by DDPG agent

TABLE IV: Summary of neural network topologies.

Case Name	# of Hidden Layers	# of Neurons	Total Parameters
IEEE 37 Bus	2	200, 300	153,210
IEEE 123 Bus	3	300, 400, 400	710,020

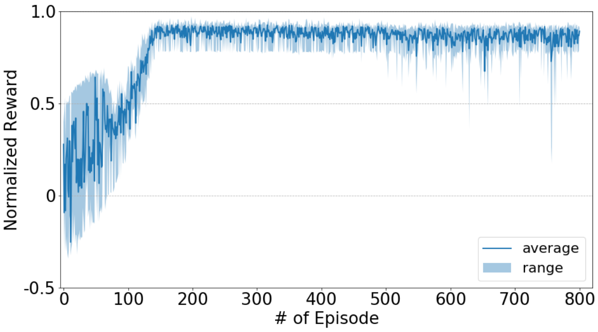

Fig. 7: Normalized reward during the training process for the IEEE 123 bus feeder. The dark blue line represents the average reward of 5 experiments with different random seeds, and the shaded light blue area displays the range of episode-averaged rewards of those 5 experiments.

C. Case study setup

The trained DDPG agent is used to perform grid voltage control for one year with 1 hour resolution (8760 different scenarios). The PV generation and load consumption profiles for the test are from public data and are plotted in Fig. 8. The 8760 test scenarios cover a wide variety of realistic operation conditions (over-voltages, under-voltages, and no voltage violation) the DDPG agent can experience if deployed online. A desirable DDPG agent needs to respond to all possible power system voltage conditions properly during online operation. While randomly generated data is used in training stage, the test stage uses real PV and load profiles and no test data is used in traing. To test the robustness of a DDPG that was trained solely on randomly generated episodes, online training is not applied. In other words, reward feedbacks after taking the suggested actions are not used to retrain and improve the DDPG agent during the test. Therefore, the DDPG agent makes decision solely based on the past experiences learned during the training phase. Contrary to the iterative process in the training stage, where the DDPG could iterate many times to reach an action with a high reward, in the test stage the agent has to provide effective actions within one iteration.

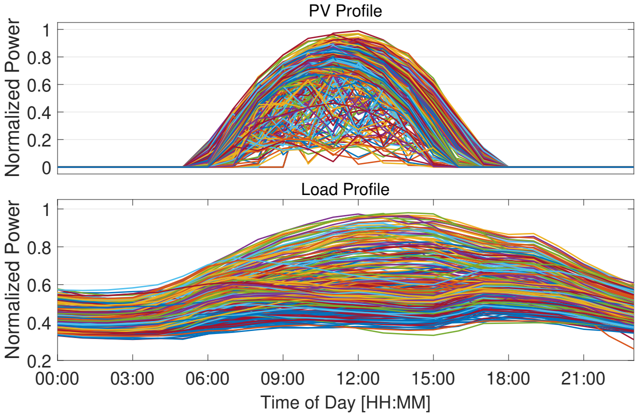

Fig. 8: Normalized PV (top) and load (bottom) profiles used for tests. The PV generation profile is from public solar power datasets maintained by NREL [48]. The load profile is from the OpenDSS installation directory [44]. The same profile is used for all PVs and loads. 1 year (365 profiles) are shown.

TABLE V: Summary of test results for IEEE 37 bus feeder. The values shown are the cumulative quantities of each parameter for 1 year.

	# of Under-voltages	# of Over-voltages	Max Volt [p.u.]	Min Volt [p.u.]	PV Curtailment (kWh)
No Control	1,797	6,998	1.067	0.921	0
Volt-Var	0	0	1.040	0.966	15,252
OPF	0	0	1.050	0.950	1,291
DDPG	0	0	1.046	0.961	1,854

V. Results and Discussion

A. IEEE 37 Bus Feeder

The node voltages of the IEEE 37 bus feeder are presented in Fig. 9. For the no control case (i.e. no reactive power generation from PVs), the feeder experiences both over-voltages and under-voltages. The total number of over-voltage violations, where one node voltage out of ANSI limits at any time step (a new scenario is applied at each time step) counts as one violation, is 6,998 with the maximum voltage equaling to 1.067 p.u. (Table V). The total under-voltage occurrence is 1,797 with a minimum voltage of 0.921 p.u.. With Volt-Var droop control (Fig. 1), all voltage violations are mitigated. The maximum voltage is brought down to 1.040 p.u. while the minimum voltage is increased to 0.966 p.u.. The OPF also eliminates all voltage violations. The maximum voltage and minimum voltage are 1.050 and 0.950 p.u., respectively, reflecting the voltage constraints imposed in (20). The proposed DDPG also solves all voltage issues, reducing the maximum voltage to 1.046 p.u. while boosting the minimum voltage to 0.961 p.u.. The results demonstrate that a DDPG agent is not only effective in voltage regulation but also robust in various operation conditions.

Reactive power generation by SIs to resolve voltage issues could lead to PV production curtailment due to the capacity limit as shown in Fig. 2. The normalized total PV curtailments due to reactive power utilization are plotted in Fig. 10. Since reactive power usage is prohibited in the no control case, the corresponding curtailment is always zero. For the Volt-Var, OPF, and DDPG cases, the active power of the SI needs to be curtailed to make room for reactive power generation. However, OPF and DDPG coordinate different SIs for voltage regulation, ensuring that reactive power is used more efficiently and less curtailment is incurred in comparison to Volt-Var. Relative to the Volt-Var case, DDPG reduces the curtailment by 88% (from Volt-Var 15,252 to DDPG 1,854 kWh) while OPF provides the optimal solution of most efficiently dispatching reactive power for voltage regulation and reduces the curtailment by 92% (from Volt-Var 15,252 to OPF 1,291 kWh). The difference in curtailment between DDPG and OPF is only 563 kWh for the 1 year test period, i.e. only 1.54 kWh per day. Therefore, DDPG approaches the optimal solution in minimizing reactive power to resolve voltage issues.

The non-linear and non-convex nature of the OPF problem renders it computationally intensive, limiting its practical application. On the contrary, a trained DDPG agent can map grid state information directly to SI actions, which requires only one feed-forward step of the neural network and is extremely efficient. Table VI compares the solution time of OPF and DDPG. The solution time is defined as the mean value of solving 8,760 different scenarios. For OPF, this is the time needed to solve the optimization problem. For DDPG, this is the time the DDPG agent takes to make decisions on SI actions after receiving grid state information. The solution time is not shown for the Volt-Var method, as it is an autonomous local control scheme that acts essentially instantaneously.

TABLE VI: Solution time comparison on the IEEE 37 bus feeder.

Method	OPF	DDPG
Solution Time (s)	27.6	1.5×10^{-3}

The simulations are carried out on a PC with Intel (R) Core(TM) i7-4700MQ 2.8-GHz processor using Python 3.7. The OPF is formulated and solved using the KNITRO solver [49] with Pyomo interface [50]. The DDPG averages only 1.5 ms (CPU time) to make decisions while the OPF needs 27.6 s (CPU time) to get the solution (Table VI).

B. IEEE 123 Bus Feeder

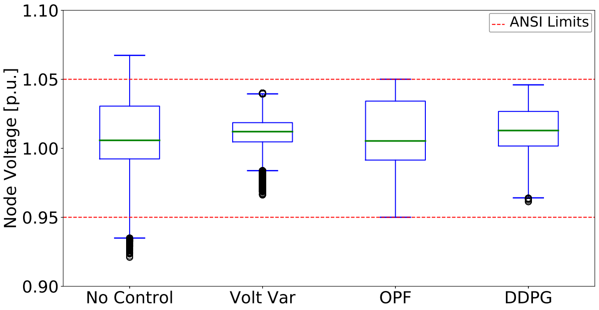

Fig. 9: Boxplot of node voltages of the 1 year test with 8760 scenarios on the IEEE 37 bus test feeder. For each case, the boxplot contains voltages of all nodes of all 8760 test scenarios (324,120 data points). The red dashed lines represent the [0.95,1.05] p.u. ANSI limits [43].

The proposed DDPG is also tested on the larger IEEE 123 bus feeder. Since the academic license of the KNITRO [49] solver is limited to 300 variables and 300 constraints, the OPF is not compared here.

Fig. 11 displays nodal voltage distributions of the three remaining cases. Numerous voltage violations can be observed for no control case with a maximum voltage of 1.066 p.u. and a minimum voltage of 0.906 p.u.. There are 19,865 over-voltage and 7,832 under-voltage occurrences for the no control case (Table VII). The autonomous Volt-Var control eliminates all over-voltages, reducing the maximum voltage to 1.040 p.u.. However, there are still 100 under-voltages (Table VII) and the minimum voltage is 0.946 p.u.. With the proposed DDPG control, all voltage violations are resolved: the maximum voltage observed is 1.0497 p.u. and the minimum voltage is 0.9503 p.u.. The DDPG agent successfully

TABLE VII: Summary of test results for IEEE 123 bus feeder. The violations and curtailment numbers are the cumulative quantities for 1 year.

	# of Under-voltages	# of Over-voltages	Max Volt [p.u.]	Min Volt [p.u.]	PV Curtailment (kWh)
No Control	7,832	19,865	1.066	0.906	0
Volt-Var	100	0	1.040	0.946	58,703
DDPG	0	0	1.0497	0.9503	7,277

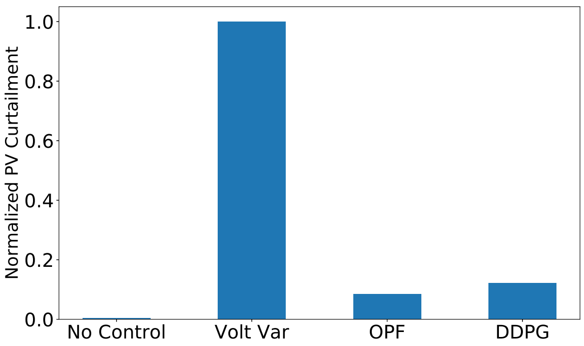

Fig. 10: Normalized total energy curtailed during the 1 year test with 8,760 scenarios for four different control cases on the IEEE 37 bus feeder. The energy curtailed is normalized with the energy curtailed of the Volt-Var case. The PV curtailment here is defined as the PV active power generation deficit between no control case (no reactive power utilization) and the other three cases (with reactive power generation), providing a direct comparison of Volt-Var, OPF, and DDPG on effective usage of SI reactive power.

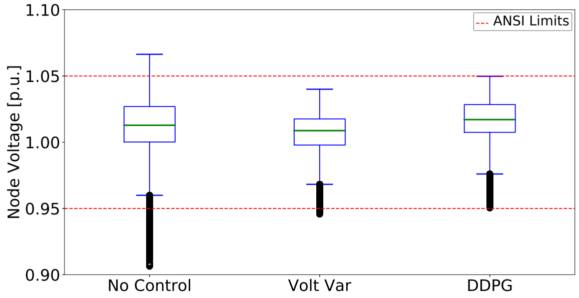

Fig. 11: Boxplot of node voltages of 1 year test with 8,760 scenarios on the IEEE 123 bus feeder. For each case, the boxplot consists of voltages of all nodes of all 8,760 scenarios (1,121,280 data points). The red dashed lines represent the [0.95,1.05] p.u. ANSI limits [43].

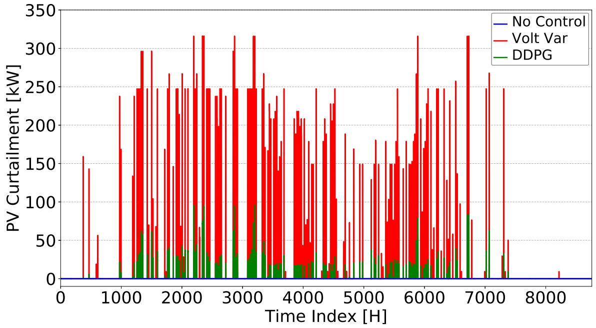

Fig. 12: Total PV generation curtailment for each scenario for the IEEE 123 bus feeder for no control (zero curtailment), Volt-Var, and DDPG.

learned a delicate strategy to utilize the minimal amount of reactive power to keep the voltage just within the ANSI limits. This demonstrates that enforcing voltage limits through large voltage violation penalties (Section III-B) is effective.

The curtailment of PV production due to reactive power utilization is displayed in Fig. 12. Since DDPG coordinates different SIs to utilize reactive power more efficiently, much less curtailment is incurred in comparison to the Volt-Var case. The total energy curtailed for the DDPG case represents a 88% reduction in curtailment (from Volt-Var 58,703 to DDPG 7,277 kWh, as shown in Table VII).

The average decision making time of the DDPG on the IEEE 123 bus feeder is 1.6×10^{-3} s (CPU time), which is almost the same as for the IEEE 37 bus feeder. This again demonstrates the solution speed advantage of the DDPG approach.

VI. Conclusions

In this paper, a DDPG-based method is proposed to coordinate multiple SIs for distribution network voltage regulation. Comprehensive tests with thousands of realistic scenarios are conducted on the IEEE 37 bus feeder and the IEEE 123 bus feeder to evaluate the trained DDPG agents. The proposed DDPG approach is compared against autonomous Volt-Var control and OPF. The results demonstrate that even without online reward feedbacks, a well-trained DDPG agent can rely solely on the knowledge accumulated in the training phase to make robust decisions under various operation conditions. As DDPG can coordinate different SIs, it is more effective in mitigating voltage issues and does so with much less reactive power compared to autonomous Volt-Var control, achieving significant reduction in PV production curtailment. The DDPG decisions are as effective as the optimal solutions from OPF in terms of resolving voltage problems; however, the DDPG results in a marginal increase in PV curtailment over OPF due to slightly more reactive power usage.

The OPF approach relies on accurate forecasting of future conditions, due to its large computation time, which would significantly increase for larger networks. While this paper assumes perfect forecasts for the OPF, forecast errors in actual applications can lead to performance deterioration for OPF including a failure to maintain ANSI voltage limits. On the contrary, (assuming fast communications), the DDPG is independent of forecasts, as it is capable of reaching decisions instantaneously.

Data Availability

The data that supports the findings of this study are available from the corresponding author upon reasonable request.

References

[1] 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. [Online]. Available: https://doi.org/10.1109/icmla.2019.00310

[2] R. Walling, R. Saint, R. C. Dugan, J. Burke, and L. A. Kojovic, "Summary of distributed resources impact on power delivery systems," *IEEE Transactions on power delivery*, vol. 23, no. 3, pp. 1636–1644, 2008.

[3] C. Li, E. Ratnam, and J. Kleissl, "Distribution feeder hotspots," California Solar Initiative RD&D Program, Tech. Rep., 2016.

[4] Z. K. Pecenak, H. V. Haghi, C. Li, M. J. Reno, V. R. Disfani, and J. Kleissl, "Aggregation of voltage-controlled devices during distribution network reduction," *IEEE Transactions on Smart Grid*, pp. 1–1, 2020. [Online]. Available: https://doi.org/10.1109/tsg.2020.3011073

[5] P. M. Carvalho, P. F. Correia, and L. A. Ferreira, "Distributed reactive power generation control for voltage rise mitigation in distribution networks," *IEEE transactions on Power Systems*, vol. 23, no. 2, pp. 766–772, 2008.

[6] Z. K. Pecenak, J. Kleissl, and V. R. Disfani, "Smart inverter impacts on california distribution feeders with increasing pv penetration: A case study," in *2017 IEEE Power & Energy Society General Meeting*. IEEE, 2017, pp. 1–5.

[7] "Ieee standard for interconnection and interoperability of distributed energy resources with associated electric power systems interfaces," *IEEE Std 1547–2018*, 2018.

[8] California Public Utilities Commission *et al.*, "Rule 21 interconnection," 2018.

[9] Hawaiian Electric, Maui Electric, and Hawaii Electric Light, "Hawaiian electric rules," 2018.

[10] T.-T. Ku, C.-H. Lin, C.-S. Chen, and C.-T. Hsu, "Coordination of transformer on-load tap changer and pv smart inverters for voltage control of distribution feeders," *IEEE Transactions on Industry Applications*, vol. 55, no. 1, pp. 256–264, 2018.

[11] H. M. Maruf and B. H. C. Energy, "Impact of smart inverter functions on dynamic step voltage regulator settings for distribution voltage control," in *2019 IEEE Milan PowerTech*. IEEE, 2019, pp. 1–6.

[12] P. P. Singh, "Coordinated voltage control in active distribution network with on-load tap changer and solar pv system systems," 2019.

[13] 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.

[14] P. Šulc, S. Backhaus, and M. Chertkov, "Optimal distributed control of reactive power via the alternating direction method of multipliers," *IEEE Transactions on Energy Conversion*, vol. 29, no. 4, pp. 968–977, 2014.

[15] 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.

[16] C. Cortés, H. V. Haghi, C. Li, and J. Kleissl, "Dynamic weight-based collaborative optimization for power grid voltage regulation," in *Proceedings of the ISES Solar World Congress 2019*. International Solar Energy Society, 2019. [Online]. Available: https://doi.org/10.18086/swc.2019.15.01

[17] C. Li, V. R. Disfani, Z. K. Pecenak, S. Mohajeryami, and J. Kleissl, "Optimal oltc voltage control scheme to enable high solar penetrations," *Electric Power Systems Research*, vol. 160, pp. 318–326, 2018.

[18] 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)*. IEEE, 2019, pp. 1–5.

[19] C. Li, V. R. Disfani, H. V. Haghi, and J. Kleissl, "Coordination of oltc and smart inverters for optimal voltage regulation of unbalanced distribution networks," *Electric Power Systems Research*, vol. 187, p. 106498, 2020.

[20] K. Christakou, J.-Y. LeBoudec, M. Paolone, and D.-C. Tomozei, "Efficient computation of sensitivity coefficients of node voltages and line currents in unbalanced radial electrical distribution networks," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 741–750, jun 2013. [Online]. Available: https://doi.org/10.1109/tsg.2012.2221751

[21] A. Borghetti, M. Bosetti, S. Grillo, S. Massucco, C. A. Nucci, M. Paolone, and F. Silvestro, "Short-term scheduling and control of active distribution systems with high penetration of renewable resources," *IEEE Systems Journal*, vol. 4, no. 3, pp. 313–322, 2010.

[22] Q. Nguyen, H. V. Padullaparti, K.-W. Lao, S. Santoso, X. Ke, and N. Samaan, "Exact optimal power dispatch in unbalanced distribution systems with high pv penetration," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 718–728, 2018.

[23] K. Christakou, M. Paolone, and A. Abur, "Voltage control in active distribution networks under uncertainty in the system model: a robust optimization approach," *IEEE Transactions on Smart Grid*, 2017.

[24] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, p. 354, 2017.

[25] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[26] R. Diao, Z. Wang, D. Shi, Q. Chang, J. Duan, and X. Zhang, "Autonomous voltage control for grid operation using deep reinforcement learning," in *2019 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, Aug. 2019. [Online]. Available: https://doi.org/10.1109/pesgm40551.2019.8973924

[27] 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 Transactions on Power Systems*, vol. 35, no. 1, pp. 814–817, 2019.

[28] J. Duan, H. Li, X. Zhang, R. Diao, B. Zhang, D. Shi, X. Lu, Z. Wang, and S. Wang, "A deep reinforcement learning based approach for optimal active power dispatch," in *2019 IEEE Sustainable Power and Energy Conference (iSPEC)*. IEEE, Nov. 2019. [Online]. Available: https://doi.org/10.1109/ispec48194.2019.8974943

[29] T. Lan, J. Duan, B. Zhang, D. Shi, Z. Wang, R. Diao, and X. Zhang, "Ai-based autonomous line flow control via topology adjustment for maximizing time-series atcs," *arXiv preprint arXiv:1911.04263*, 2019.

[30] 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 Transactions on Power Systems*, pp. 1–1, 2020. [Online]. Available: https://doi.org/10.1109/tpwrs.2020.2990179

[31] 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, 2019.

[32] H. Xu, A. D. Domínguez-García, and P. W. Sauer, "Optimal tap setting of voltage regulation transformers using batch reinforcement learning," *arXiv preprint arXiv:1807.10997*, 2018.

[33] W. Wang, N. Yu, J. Shi, and Y. Gao, "Volt-var control in power distribution systems with deep reinforcement learning," *IEEE SmartGridComm*, 2019.

[34] 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, Jul. 2020. [Online]. Available: https://doi.org/10.1109/tsg.2019.2962625

[35] Y. Gao, W. Wang, and N. Yu, "Consensus multi-agent reinforcement learning for volt-var control in power distribution networks," *arXiv preprint arXiv:2007.02991*, 2020.

[36] R. E. Helou, D. Kalathil, and L. Xie, "Fully decentralized reinforcement learning-based control of photovoltaics in distribution grids for joint provision of real and reactive power," *arXiv preprint arXiv:2008.01231*, 2020.

[37] H. Liu and W. Wu, "Online multi-agent reinforcement learning for decentralized inverter-based volt-var control," *arXiv preprint arXiv:2006.12841*, 2020.

[38] D. Cao, J. Zhao, W. Hu, F. Ding, Q. Huang, and Z. Chen, "Distributed voltage regulation of active distribution system based on enhanced multi-agent deep reinforcement learning," *arXiv preprint arXiv:2006.00546*, 2020.

[39] 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, 2004.

[40] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[41] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," 2014.

[42] 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.

[43] "American national standard for electric power systems and equipment-voltage ratings (60 hertz)," *ANSI std. C84. 1-2011*, 2011.

[44] R. C. Dugan, "The open distribution system simulator (opendss)," *EPRI OpenDSS Manual*, 2012.

[45] V. Rigoni and A. Keane, "Open-dsopf: an open-source optimal power flow formulation integrated with opendss," in *2020 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, Aug. 2020.

[46] N. Van Der Borg and A. Burgers, "Inverter undersizing in pv systems," in *3rd World Conference onPhotovoltaic Energy Conversion, 2003. Proceedings of*, vol. 2. IEEE, 2003, pp. 2066–2069.

[47] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," *arXiv preprint arXiv:1604.02201*, 2016.

[48] GE Energy, "Western wind and solar integration study," No. NREL/SR-550-47434. National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2010.

[49] R. H. Byrd, J. Nocedal, and R. A. Waltz, "Knitro: An integrated package for nonlinear optimization," in *Large-scale nonlinear optimization*. Springer, 2006, pp. 35–59.

[50] W. E. Hart, C. D. Laird, J.-P. Watson, D. L. Woodruff, G. A. Hackebeil, B. L. Nicholson, and J. D. Siirola, *Pyomo-optimization modeling in python*. Springer, 2017, vol. 67.