Centralized Coordination of DER Smart Inverters using Deep Reinforcement Learning

Daniel Glover, *Graduate Student Member, IEEE*, Anamika Dubey, *Senior Member, IEEE*

***Abstract*—The modern power grid continues to grow in complexity and dynamics due to the addition of various inverter based resources (IBRs), which require further oversight from system operators. Maintaining adequate system-wide voltage regulation through remote control of distributed solar photovoltaic (PV) inverters offers flexibility for grid operators, but becomes computationally challenging when using traditional optimization approaches due to model complexities, measurement uncertainties, and increasing numbers of interconnected devices. This work proposes a Deep Reinforcement Learning (DRL) based model-free volt-var control (VVC) of smart inverters for an optimal system-wide voltage regulation. Our controller also includes a limited set of reactive power dispatch rules for smart inverters, specified by the IEEE 1547-2018 standard, through an informative reward design process to embed these additional operational constraints. Preliminary results carried out on the IEEE 123 bus network demonstrate the ability of a single centralized controller with no prior system knowledge to achieve successful bus voltage deviation minimization through VVC selective action learning of several solar PV units. Furthermore, the results show the effectiveness of incorporating learnable barrier limits within the reward function design, and expose the importance of reward shaping and variability in DRL algorithms.**

2023 IEEE Industry Applications Society Annual Meeting (IAS) | 979-8-3503-2016-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IAS54024.2023.10406599

I. Introduction

A. Motivation for Centralized Learning-Based Control

Traditional distribution networks are shifting from passive delivery systems to active, bi-directional networks hosting a mixture of legacy and smart devices [1]. Increasing penetrations of distributed energy resources (DERs) introduces various uncertainties into system operations, reducing the accuracy of deterministic solutions, and requiring attention to faster timescale dynamics. The simultaneous deployment of measurement devices, both at the grid-edge (via advanced metering infrastructures) and at the system-level (using microPMUs and other wide-area monitoring systems) provide massive amount of potential real-time data that can be used for operational decision-making [2].

With the influx of DERs at the grid edge, optimal Volt-Var control (VVC) has been extensively studied for the goal of providing reactive power support to enhance feeder voltages and achieve other operational benefits such as conservation voltage reduction or loss minimization, etc. Moreover, to address the voltage regulation problem due to growing penetrations of DERs, IEEE 1547-2018 standard was established to allow for reactive power support via local autonomous control of smart inverters [3]. From the system operations standpoint, the coordination among smart inverters is vital to meet global system-level objectives. Various traditional optimization and control approaches have been developed in past; however, due to various uncertainties resulting from measurement and model errors, and the problem dimentionality, traditional optimization-based models are extremely difficult to solve within the desired operational time-scale [4], [5].

For example, centralized and distributed optimal power flow (OPF) methods have been developed to determine optimal setpoint dispatch for DERs to minimize power losses and limit operating limit violations while maximizing power delivery to the consumer. The non-linearity of the OPF problems is generally handled using several convex relaxation techniques [6]. However, due to forecasting and physical model uncertainties, existing models may only provide limited information or unsatisfactory input data resulting in poor quality solutions [7]. The optimization under uncertainty requires stochastic or robust optimization approaches, both are extremely difficult to scale and generalize for general noise distributions. For example, stochastic programming formulations under adverse uncertainty for unit commitment are discussed in [8], [9] where robust non-linear formulations require longer time to achieve an optimal solution and can be difficult to scale.

Another complexity for inverter control arises from need to incorporate difficult mathematical constraints resulting from the IEEE 1547-2018 standard. For example, in [10], authors use distributed OPF to learn VVC and Volt-Watt control (VWC) droop control settings by incorporating piecewise curves with the addition of the inverter standards. However, embedding the IEEE 1547-2018 standard into the optimization problem introduces integer variables, and the non-linear power flow model is simplified to form a Mixed Integer Linear Programming (MILP) problem that proves more difficult to solve due to the introduction of larger numbers of additional integer variables. Similarly, [11] formulates centralized VVC for DERs to minimize operational cost as a Mixed Integer Nonlinear Program (MINLP), that are extremely difficult to solve or scale. Looking ahead, it is estimated that these pitfalls will become increasingly problematic as the grid becomes more complex, motivating data driven control methods that are robust to uncertainties, which can learn and make decisions in diverse environments.

B. Deep Reinforcement Learning for Volt-Var Control

Reinforcement learning (RL) for power systems applications have been proven to provide strong performance under model and measurement uncertainties when compared to traditional optimization and control approaches [12]. However, they suffer from lack of exploration and safety guarantees, scalability issues and policy degradation over time leading to poor online transfer, known as the *simulation-to-reality* gap [13]. In [14], authors discussed the use of constrained optimization techniques in deep RL (DRL) problems for safety, that often involve setting fixed limits on exploration boundaries; however, they do not discuss learning these parameters as formulaic reward objectives. It has been shown that fixing bounds of the observation and state spaces to enforce strict exploration rules can lead to online under-performance, conservatism, or potential failure when exposed to different/varying real world conditions/scenarios [15].

Specifically on VVC using DRL, [16] establishes a DRL-based VVC to minimize operational cost while respecting physical constraints using Trust Region Policy Optimization (TRPO); however, they do not learn the physical boundaries of the system and address scalability in larger networks. In [17], authors use a Soft Actor Critic (SAC) algorithm for coordinated inverter control via Markov Game and experience replay buffer. The method achieves promising results for reduction of power curtailment, but includes hard constraints of inverter regulations in the OPF model as opposed to the DRL learning process, which could lead to online control exploration actions that may violate limits when exposed to a different data distribution. Finally, [18] uses Deep Deterministic Policy Gradient (DDPG) with Actor Critic model for VVC, but uses a reward function with large penalty resulting in potential scalability and *sim-to-real* policy transfer challenges, or failure over time when exposed to unseen operational conditions.

The contribution(s) of this work is to demonstrate the capabilities of a centralized DRL agent learning VVC for a distribution network (with no prior system knowledge) containing randomly distributed DERs through informative reward design of physical solar inverter regulatory standards and active action selection. Our simulation results using an Advantage Actor-Critic (A2C) algorithm to learn VVC controls are promising in achieving system-level goals to reduce nodal voltage deviations and power losses while learning to operate within regulatory standards as specified by the existing IEEE 1547-2018 standard [3]. Note that IEEE 1547-2018 standard is fairly comprehensive and in this paper, we include only a limited set of inverter operating conditions (Section 5.2 in [3]) that previously have not been included in the related literature. Our future work will entail more comprehensive treatment on this subject via including recommended Volt-VAR and Volt-WATT curves. This study also aims to expose limitations of the learning algorithm to emphasize further study on obtaining safe solutions for model-free DRL controllers.

II. Problem Formulation

This section details a systematic approach to develop the proposed DRL-based VVC controller. First, we introduce a mathematical formulation for the centralized OPF-based VVC with the goal of dispatching reactive power setpoints for solar PVs with smart inverters, referred to as DERs. The distribution system model with the VVC objective is discussed, along with the device model for DERs. The optimization problem is used to motivate the DRL model for VVC formulated as a constrained *Markov Decision Process* (MDP). To fully describe the DRL model, we introduce the state and action space definitions and multi-objective reward design. We also introduce the *Advantage Actor-Critic* (A2C) algorithm as model-free approach to train DRL-based VVC controller.

A. Centralized OPF and Distribution System Model

We assume a distribution network graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$ containing a set of $\mathcal{N}$ nodes and $\mathcal{E}$ lines (edges) such that $\mathcal{N} = \{1 : N\}$, where DERs have been installed at $\mathcal{N}_{\mathcal{DER}}$ buses such that $|\mathcal{N}_{\mathcal{DER}}| \leq |\mathcal{N}|$ and $\mathcal{N}_{\mathcal{DER}} \subset \mathcal{N}$. Any two nodes i and j are connected by an edge (i, j) representing a physical line connection, where node i is the parent of child node j. The objective of the central controller is to dispatch optimal reactive power setpoints to selected DERs in the system based on the measured voltages at the DER buses, which are centrally collected at each time step t; t is omitted in the formulation for brevity.

The objective function in (1) minimizes voltage deviations $V_{dev} = \sum_{i=1}^{N}(v_i - 1)^2$ from 1 pu at all buses under observation, total system active power losses $P_{loss} = \sum P_{gen} - \sum P_{load}$, and provides a metric Q^{ρ}_{DER} to indicate a violation of VVC limits from those stated in IEEE 1547-2018 [3] (see next section). The VVC must also abide by the physical, operational and technical constraints of the power system and its components (1a) - (3).

$$min \sum_{i=1}^{N} V_{dev} + P_{loss} + Q_{\rho_{DER}} \tag{1}$$

subject to

$$P_i^{inj} - P_i^{spec} = V_i \sum_{j=1}^{N} V_j Y_{ij} \cos(\delta_i - \delta_j - \theta_{ij}) - (P_{DER,i} - P_{D,i}) \tag{1a}$$

$$Q_i^{inj} - Q_i^{spec} = V_i \sum_{j=1}^{N} V_j Y_{ij} \sin(\delta_i - \delta_j - \theta_{ij}) - (Q_{DER,i} - Q_{D,i}) \tag{1b}$$

$$P_{ij}^{min} \leq P_{ij} \leq P_{ij}^{max}, \quad \forall (i,j) \in \mathcal{E} \tag{1c}$$

$$Q_{ij}^{min} \leq Q_{ij} \leq Q_{ij}^{max}, \quad \forall (i,j) \in \mathcal{E} \tag{1d}$$

Constraints (1a)-(1d) delineate the mismatch between real and reactive injected power P_i^{inj}, Q_i^{inj} at a bus i and the specified powers P_i^{spec}, Q_i^{spec}, equivocal to generation minus demand $P_{DER,i} - P_{D,i}, Q_{DER,i} - Q_{D,i}$, with line flow limits on P_{ij}, Q_{ij}. Similarly, the operational voltage limits for all buses as per the ANSI Standard [19] are given as $V_i^{max} = 1.05$ pu and $V_i^{min} = 0.95$ pu in (2). Finally, each DER is limited by its rated apparent power, $S_{DER,i}^{Rated}$ in (3).

$$V_i^{min} \leq V_i \leq V_i^{max}, \quad \forall i \in \mathcal{N} \tag{2}$$

$$0 \leq \sqrt{P_{DER,i}^2 + Q_{DER,i}^2} \leq S_{DER,i}^{rated}, \quad \forall N_{DER} \in \mathcal{N} \tag{3}$$

There are additional constraints imposed on reactive power dispatch from DER ($Q_{DER,i}$) as per the recommendations from IEEE 1547-2018 standard [3]. A limited set of operational constraints on inverter reactive power limits as per

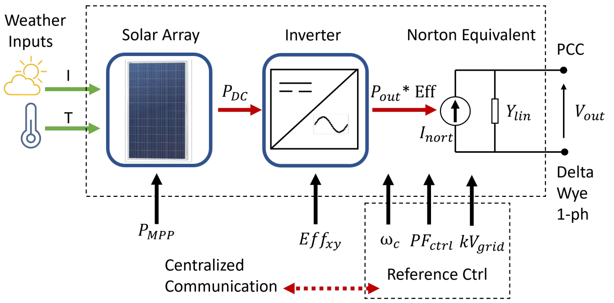

Figure 1. PV System Single Phase Model

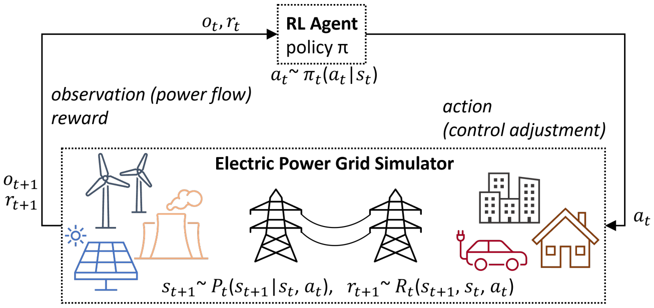
o_t, r_t
RL Agent
policy π
$a_t \sim \pi_t(a_t|s_t)$
observation (power flow)
reward
action
(control adjustment)
Electric Power Grid Simulator
o_{t+1}
r_{t+1}
a_t
$s_{t+1} \sim P_t(s_{t+1}|s_t, a_t)$, $r_{t+1} \sim R_t(s_{t+1}, s_t, a_t)$

Figure 2. RL agent power system interaction process

the IEEE 1547-2018 standard, especially under low lighting conditions, are included in this paper, further discussed in the following section.

B. DER Model: PV with Smart Inverter Control

The DER model, used in this paper, is shown in Figure 1. It combines the solar array panel with inputs temperature T and irradiance I, maximum power point tracking (MPPT) as P_{MPP} maximum real power delivery at unity power factor, and inverter module with efficiency curves Eff_{xy} provided from [20]. The output is represented as a Norton equivalent current source I_{nort} at the point of common coupling (PCC). Each PV system includes grid reference control monitoring for frequency and local grid voltage . We also assume a physical line of communication exists for centralized monitoring and control as per grid requirements in [3].

The recommended ratings for smart inverter requires having 44% of $S^{rated}_{DER,i}$ being available when PV is generating its peak active power $P^{rated}_{DER,i}$. However, additional operational limits are imposed as stated in (4a) - (4b). These are defined as per Section 5.2 of the IEEE 1547-2018 standard for voltage reactive power control of category A and B inverters [3]. Specifically, the available reactive power absorptions, $Q^{limit}_{DER,i}$ and injections are limited to 44% of $S^{rated}_{DER,i}$, but also depend on the amount of active power produced as a percentage of $S^{rated}_{DER,i}$. During low sunlight hours, $Q^{limit}_{DER,i}$ decreases further in (4a) as a fraction of the actual real power being produced by 20% of the inverter rated real power $P^{rated}_{DER,i}$.

$$Q^{limit}_{DER,i} \leq 0.44 S^{rated}_{DER,i} \times \frac{P_{DER,i}}{0.2 \times P^{rated}_{DER,i}}, \quad \text{if} \;\; P_{DER,i} < 0.2 S^{rated}_{DER,i} \tag{4a}$$

$$0.44 S^{rated}_{DER,i} \leq Q^{limit}_{DER,i} \leq \sqrt{((S^{rated}_{DER,i})^2 - P^2_{DER,i})} \tag{4b}$$

else if, $P_{DER,i} \geq 0.2 S^{rated}_{DER,i}$.

We would like to highlight that there are additional and more comprehensive set of recommendation on inverter reactive control as per the IEEE 1547-2018, such as the recommended Volt-VAR curves included in [10]. We aim to include these difficult constraints as an extension of this work.

C. Constrained MDP Formulation for Volt-Var Dispatch

This section describes the formulation of centralized VVC DER dispatch as a constrained Markov Decision Process (CMDP). The sequential nature of the RL learning process can be modeled as an MDP defined by the tuple $< \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \mathcal{C} >$, consisting of an action space $\mathcal{A}$, state space $\mathcal{S}$, the transition probability function $P(s_t, a_t, s'_t) = Pr(s'_t|s_t, a_t) = \mathcal{T} : \mathcal{S}X\mathcal{A} \to \mathcal{S}'$, reward function $r(s_t, a_t) : \mathcal{S}X\mathcal{A} \to \mathbb{R}$, and $\mathcal{C}$, a set of constraints applied to the learned policy $\pi(a_t|s_t) \in \mathcal{A}$. Generally speaking, the environment begins in an initial state $s_0 \in \mathcal{S}$, and at each time step t, the DSO (distribution system operator) agent chooses action $a_t \in \mathcal{A}$ and receives a reward $r(s_t, a_t)$ dependent on the current state/action pair, after which the system moves to the next state s_{t+1} generated from $P(\cdot|s_t, a_t)$ (see Figure 2).

The goal of the central controller is to find the optimal policy π^* (5) for dispatching VVC setpoint adjustments to maximize the expected value of the discounted reward over a given time horizon. In the model-free approach taken due to various uncertainties, π^* is derived without explicitly learning the model when it becomes difficult to learn or express using *Q-Learning*. Instead of learning the optimal value function, the optimal Q-function Q^* is learned directly as $Q(s, a)$ given in (6) for an infinite time horizon starting from an initial state and action. A discount parameter γ [0:1] is used to determine the value of future rewards.

$$\pi^* = \mathbb{E}\left[\gamma^t r(s_t, a_t)\right] \tag{5}$$

$$Q^*(s, a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s, a_0 = a\right] \tag{6}$$

1) State Space: : The observation space of each environmental state of the agent is composed of DERs and the voltages v_i measured at each respective PCC at each time step t at the respective buses. In addition, the real, reactive, and complex powers of the DER units are measured to compute the $Q^{limit}_{DER,i}$ for VVC in (7), where $Q_{DER,i} = \sqrt{S^2_{DER,i} - P^2_{DER,i}}$ for each DER.

$$Q^{\rho}_{DER,i} = |\frac{Q_{DER,i}}{Q^{limit}_{DER,i}}| \tag{7}$$

Thus, at any given time step t, the state s_t of the environment observed by the central controller is given by $s_t = [N_{DER}, v_i, P_{DER,i}, Q_{DER,i}, S_{DER,i}, Q^{\rho}_{DER,i}] \in \mathcal{S}$.

2) Action Space: : For this particular study, a unique action selection space A is constructed using a set of 1-D vectors which contain affiliated PV System IDs $\overline{PV}$ and an action selection vector $\overline{ASV}$ responsible for reactive power setpoint computation. Let the action space A be defined as a set of 1-D vectors such that at each time step t, an action $a_t \in A$ is selected from each action vector in sequence, given the possible action set in the current observed state s_t.

At each step, the agent selects an action tuple from the vector set. During observation of the PCC voltages and inverter powers, an action is selected for the *amount* of reactive power adjustment Q_{adj} constrained to (4a) and (4b) as $Q_{adj} = |(v_i - 1)| * 100 \pm Q_{csp}$ using the directly proportional relationship which exists between reactive power and voltage. The parameter Q_{adj} is computed as the percentage maximum limit of adjustment during action selection relative to the observed normalized bus/PCC voltage and the current reactive power setpoint of the inverter under examination Q_{csp}. In this manner, the agent is trained to make more conservative adjustments when the local grid health is optimal.

3) Multi-Objective Reward Function:: The multi-objective reward function uses a negative penalty scheme to train the agent on VVC . First, a voltage deviation penalty utilizes local voltage measurement v_i compared to the target 1 pu, with total system loss given as $P_{loss} = \sum P_{gen} - \sum P_{load}$. Finally, $Q^{\rho}_{DER,i} < 1$ if no 1547 violation occurs, bound by $Q^{limit}_{DER,i}$ in (4a) and (4b). The constrained objective in (8) is to maximize the expected cumulative total reward where the ith constraint must be satisfied by π in the set of allowable policies Π. However, it is also critical to incorporate system knowledge and optimization constraints into reward functions to guide the agent towards state boundary learning. More importantly, it has been shown in [21], [22] that treating constraints as learnable parameters, as opposed to fixing boundaries to limit an agent's exploration space in training, can produce more robust policies which generalize to those constraints naturally.

$$\underset{\pi\in\Pi}{max} E_{\pi} R(s_t, a_t) = -\sum_{i\epsilon N}(v_i - 1)^2 - P_{loss} - Q^{\rho}_{DER,i} \qquad (8)$$

D. Advantage Actor-Critic

Advantage Actor-Critic (A2C) is an on-policy synchronous variation of A3C (Asynchronous Advantage Actor-Critic), a deterministic multi-worker DRL algorithm which uses a *policy gradient* method averaged by all actors (bootstrapping) to make a decision. From [23], the critic estimates the value function $V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R^t|\pi, s_t\right]$ by approximating the Q-function from (6), while the actor(s) update the policy distribution $\pi(a_t|s_t, \theta)$ suggested by the critic(s) via the gradient. The critic makes use of the *Advantage* function to update the temporal difference target of the expected reward minus average reward based on the current action, given as $Adv_\pi(s, a) = r(s_t, a_t) + \gamma^t V_\pi(s_t) - V_\pi(s'_t)$ in Figure 3. Actor-Critic methods have shown tremendous success in DRL applications with scarce data, but are known for policy variability and scalability issues (discussed further in section III).

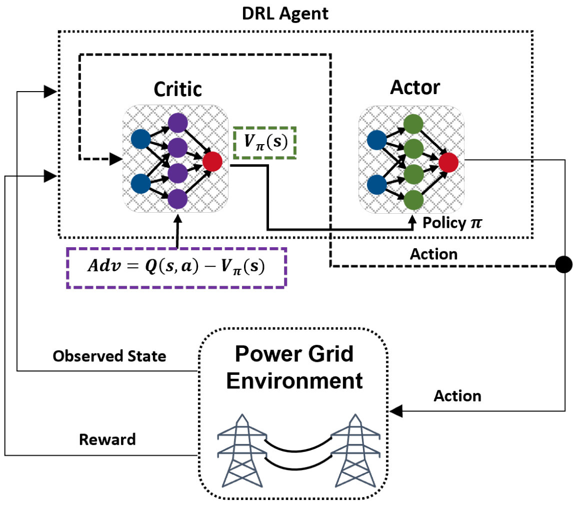

Figure 3. Advantage Actor-Critic (A2C)

III. Simulations and Results

A. Simulation Case Study

The IEEE 123 bus network used in this case study shown in Figure 4 is modified with ten solar PV systems installed at bus locations which exhibited weaker voltage profiles identified after an initial steady-state power flow. All default network voltage regulation controls are set to disengage at the initial step of each training episode, and the network is overloaded at 135% to simulate a weakened but flexible system. Each PV system is sized using maximum local nodal demand for peak PV hosting capacity $S^{Rated}_{DER,i}$. The three-phase inverter model used in this study follows the single-phase inverter equivalent circuit shown in Figure 1 (representing only a single phase), determined again by load type. The reactive power control are disabled for all PVs at initialization to allow the A2C agent to adjust inverter reactive power injections Q_{inj} or absorptions Q_{abs} via controlled setpoint adjustments. The dispatched Q targets are based on the PV system reference readings immediately following a centralized load flow, performed by the DSS power flow solver at every time step t.

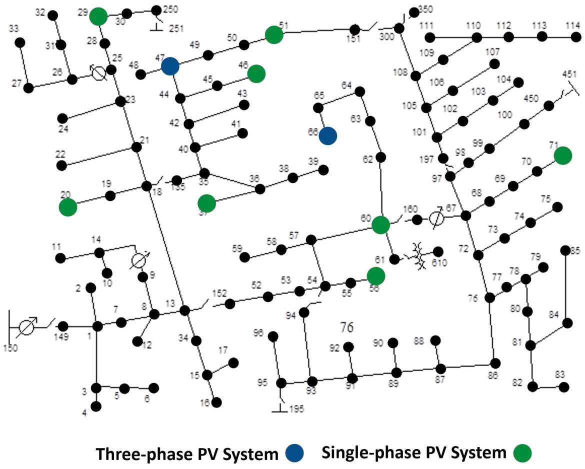

Figure 4. IEEE 123 Bus Network with Distributed Solar PV Systems

The following assumptions are also made from from IEEE1547-2018, Section V [3].

1) All inverters assume a continuous operational region of $0.88V_{Nominal} \leq V_{ref} \leq 1.1V_{Nominal}$, where $V_{Nomimal} = 1$pu and V_{ref} is the noisy measurement taken from the reference controller of the inverter
2) No fixed power factor for inverter operation is specified
3) Single line of centralized communication for remote monitoring/control exists at each PV location

Irradiance, temperature, and efficiency curves applied to each PV system timeseries simulation are taken from the National Solar Radiation Database (NSRDB) [24]. Dynamic loadshape curves from [20] are interpolated to match the time series simulation step for each load type (residential or commercial) system wide. The learning process is constructed over a consecutive 90 day finite time horizon (March - May) with a step/observation/action process frequency of 3 minutes, yielding a total of 43200 steps per training episode.

B. Simulation Results

All simulations are performed using a developed Python wrapper which combines the OpenDSS distribution simulator with *OpenAI* and DRL algorithms from *Stable Baselines* into a realistic distribution system trainable environment using *Python 3.9* on a Linux server with 24 CPU cores and 64 GB of memory.

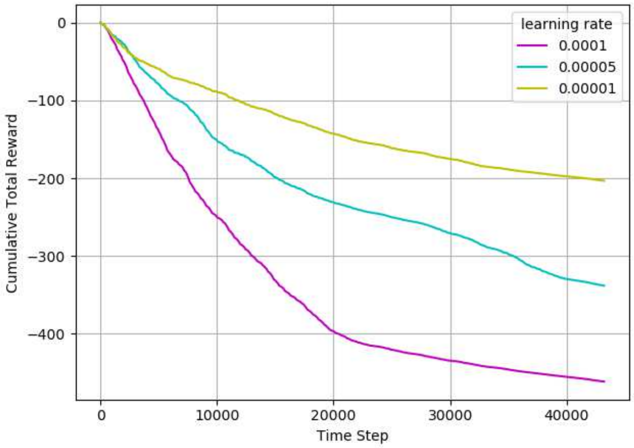

Figure 5. A2C Cumulative Rewards per Learning Rate

Figures 5 and 6 show the differences between three separate learning rates tested with the A2C algorithm for model free voltage reactive power control across all ten solar PV inverters in the network. Each learning rate converges to indicate successfully learned control policies. Figure 5 shows that the smallest learning rate of 0.00001 achieved the highest cumulative reward per episode at -200, while a learning rate of 0.00005 provided the least amount of policy variance after the task is learned. This can be seen in Figure 6 towards the last 50 episodes of training as the learning rate of 0.0001 continues to exhibit higher variance of the learned policy, while the others remain stable. Thus, the learning rate of 0.00001 is selected for the 123-bus case study.

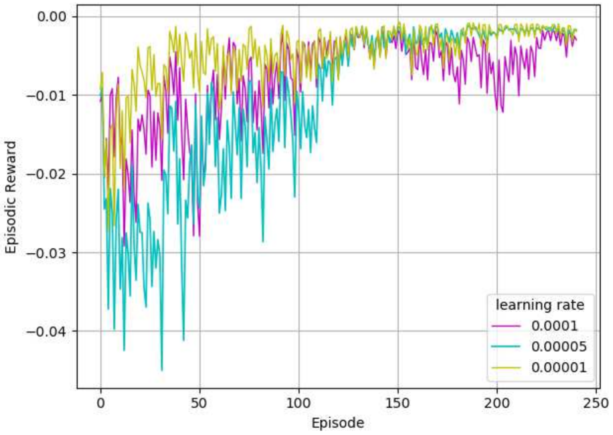

Figure 6. A2C Mean Episodic Rewards per Learning Rate

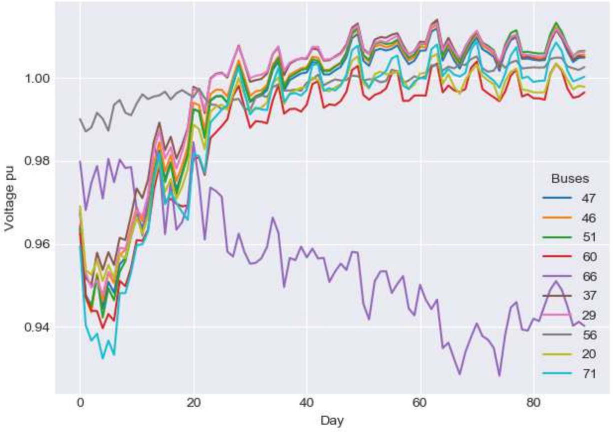

Figure 7. 90 Day Bus Voltage Profile Comparison

Figure 7 demonstrates the effectiveness of the DRL agent at coordinating bus voltage regulation among DERs via VVC dispatch. Nearly all bus voltages improved over the 90 day horizon by an average of 0.945 pu to 1.01 pu. Although a majority of nodes began the simulation at severe undervoltage levels due to heavy overloading, the central agent control correctly helped the system voltages to recover back to healthy levels. More importantly, it is noted in Figure 7 that the voltage at bus 66 was seemingly ignored from the agent's selection process during training after day 20, as seen by its voltage dropping below regulatory limits without recovery. This issue raises an interesting variability concern about the potential risks of a data-driven single centralized controller with device selection suffering from monolithic reward design.

Finally, bus voltage and solar PV power for buses 37 and 47 are compared in Figures 8 and 9, showing similar bus voltage improvement (green), with reactive power (pink) and active power (blue) output. In both cases, reactive power from both inverters remains tolerant of nameplate ratings as π^* is learned. In Figure 8, injected reactive power (neg) increases beyond regulatory limits immediately to compensate for the weakened bus voltage, but is slowly corrected over time by reward function design to achieve 1547 compliance as the policy improves. Figure 9 shows an initial absorption of reactive power by the inverter, followed by gradual injection of reactive power over time, while active power remains consistent.

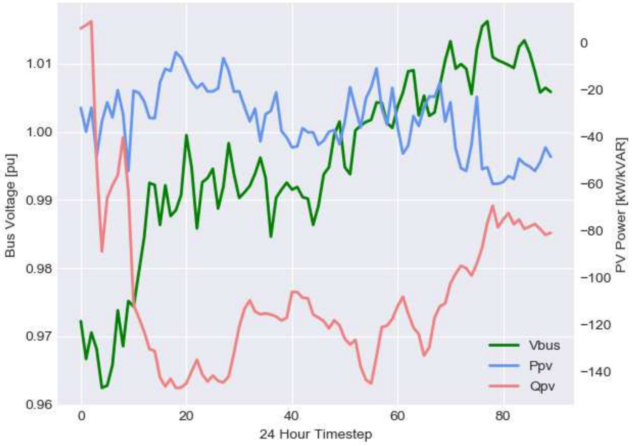

Figure 8. Bus 37 PV Voltage and Power Comparison

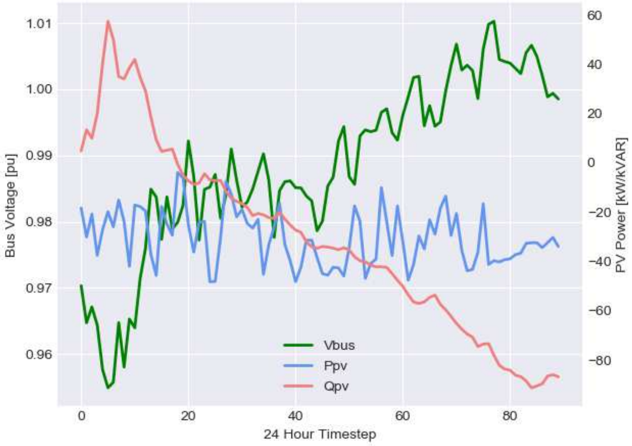

Figure 9. Bus 47 PV Voltage and Power Comparison

IV. Conclusions and Future Work Direction

In conclusion, this study shows how a DRL controller can provide centralized VVC to multiple DERs at the grid-edge, while effectively reducing system losses and learning to operate within the regulatory standards put in place for these devices. Development of a reward function which captures the optimization criterion and allows for constraint-based learning of physical rules for the power grid is essential to closing the *Sim-to-Real* gap that exists for data-driven controllers in power systems. Although control was achieved, policy variability issues prevented complete learning of all devices, as the agent differently prioritizes system objectives. Therefore, we continue to explore alternative techniques in reward shaping to enhance the robustness and reliability of DRL controllers. Future work includes the addition of a comprehensive set of volt-var curves as per [3] which will result in numerous additional hard constraints for improved safety training.

References

[1] A. Ghosal and M. Conti, “Key management systems for smart grid advanced metering infrastructure: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2831–2848, 2019.

[2] A. M. Annaswamy and M. Amin, “Ieee vision for smart grid controls: 2030 and beyond,” *IEEE Vision for Smart Grid Controls: 2030 and Beyond*, pp. 1–168, 2013.

[3] “IEEE Standard for Interconnection and Interoperability of Distributed Energy Resources with Associated Electric Power Systems Interfaces,” *IEEE Std 1547-2018*, pp. 1–138, 2018.

[4] L. Liyanarachchi, N. Hosseinzadeh, A. Mahmud, and A. Gargoom, “Challenges in power system strength assessment with inverter-based resources,” in *2021 3rd International Conference on Smart Power & Internet Energy Systems (SPIES)*. IEEE, 2021, pp. 158–163.

[5] J. Matevosyan, J. MacDowell, N. Miller, B. Badrzadeh, D. Ramasubramanian, A. Isaacs, R. Quint, E. Quitmann, R. Pfeiffer, H. Urdal *et al.*, “A future with inverter-based resources: Finding strength from traditional weakness,” *IEEE Power and Energy Magazine*, vol. 19, no. 6, pp. 18–28, 2021.

[6] M. Usman, A. Cervi, M. Coppo, F. Bignucolo, and R. Turri, “Centralized opf in unbalanced multi-phase neutral equipped distribution networks hosting zip loads,” *IEEE Access*, vol. 7, pp. 177 890–177 908, 2019.

[7] A. Bernstein and E. Dall’Anese, “Real-time feedback-based optimization of distribution grids: A unified approach,” *IEEE Transactions on Control of Network Systems*, vol. 6, no. 3, pp. 1197–1209, 2019.

[8] A. İ. Mahmutoğulları, S. Ahmed, Ö. Çavuş, and M. S. Aktürk, “The value of multi-stage stochastic programming in risk-averse unit commitment under uncertainty,” *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 3667–3676, 2019.

[9] L. Zéphyr and C. L. Anderson, “Stochastic dynamic programming approach to managing power system uncertainty with distributed storage,” *Computational Management Science*, vol. 15, no. 1, pp. 87–110, 2018.

[10] A. Inaolaji, A. Savasci, and S. Paudyal, “Distribution grid optimal power flow in unbalanced multiphase networks with volt-var and volt-watt droop settings of smart inverters,” *IEEE Transactions on Industry Applications*, vol. 58, no. 5, pp. 5832–5843, 2022.

[11] M. Tahir, R. A. El Shatshat, and M. Salama, “Reactive power dispatch of inverter-based renewable distributed generation for optimal feeder operation,” in *2018 IEEE Electrical Power and Energy Conference (EPEC)*, 2018, pp. 1–6.

[12] X. Chen, G. Qu, Y. Tang, S. Low, and N. Li, “Reinforcement learning for selective key applications in power systems: Recent advances and future challenges,” *IEEE Transactions on Smart Grid*, 2022.

[13] E. Salvato, G. Fenu, E. Medvet, and F. A. Pellegrino, “Crossing the reality gap: a survey on sim-to-real transferability of robot controllers in reinforcement learning,” *IEEE Access*, 2021.

[14] J. Garcıa and F. Fernández, “A comprehensive survey on safe reinforcement learning,” *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[15] R. Dobbe, P. Hidalgo-Gonzalez, S. Karagiannopoulos, R. Henriquez-Auba, G. Hug, D. S. Callaway, and C. J. Tomlin, “Learning to control in power systems: Design and analysis guidelines for concrete safety problems,” *Electric Power Systems Research*, vol. 189, p. 106615, 2020.

[16] W. Wang, N. Yu, J. Shi, and Y. Gao, “Volt-var control in power distribution systems with deep reinforcement learning,” in *2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 2019, pp. 1–7.

[17] Y. Pei, Y. Yao, J. Zhao, F. Ding, and K. Ye, “Data-driven distribution system coordinated pv inverter control using deep reinforcement learning,” in *2021 IEEE Sustainable Power and Energy Conference (iSPEC)*, 2021, pp. 781–786.

[18] K. Beyer, R. Beckmann, S. Geißendörfer, K. von Maydell, and C. Agert, “Adaptive online-learning volt-var control for smart inverters using deep reinforcement learning,” *Energies*, vol. 14, no. 7, p. 1991, 2021.

[19] N. E. M. Association *et al.*, *American National Standard for Electric Power Systems and Equipment-Voltage Ratings (60 Hertz)*. National Electrical Manufacturers Association, 1996.

[20] R. C. Dugan and T. E. McDermott, “An open source platform for collaborating on smart grid research,” in *2011 IEEE power and energy society general meeting*. IEEE, 2011, pp. 1–7.

[21] M. Pecka and T. Svoboda, “Safe exploration techniques for reinforcement learning–an overview,” in *International Workshop on Modelling and Simulation for Autonomous Systems*. Springer, 2014, pp. 357–375.

[22] A. Ray, J. Achiam, and D. Amodei, “Benchmarking safe exploration in deep reinforcement learning,” *arXiv preprint arXiv:1910.01708*, vol. 7, p. 1, 2019.

[23] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, “Asynchronous methods for deep reinforcement learning,” in *International conference on machine learning*. PMLR, 2016, pp. 1928–1937.

[24] M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, “The national solar radiation data base (nsrdb),” *Renewable and sustainable energy reviews*, vol. 89, pp. 51–60, 2018.