Deep Reinforcement Learning Based Approach for Optimal Power Flow of Microgrid with Grid Services Implementation

Jingping Nie, Yanchen Liu, Liwei Zhou, Xiaofan Jiang, and Matthias Preindl
Department of Electrical Engineering, Columbia University in the City of New York

***Abstract*—Electric vehicles (EVs) have rapidly grown in popularity, and the number of inverter-based EV chargers increases promptly due to their high efficiency and capabilities of providing grid services. EV and other distributed energy resources (DER) would become a crucial part of the resilience and performance of the microgrid. Optimizing the EV-interfaced microgrid is challenging due to the non-linearity and uncertainty. In this paper, we propose a method based on deep reinforcement learning (DRL) with Twin Delayed Deep Deterministic Policy Gradients (TD3) to optimize the microgrid. The proposed method can be used to optimize different objectives. An example objective of stabilizing the voltage fluctuations in a power system modified from the IEEE 30-bus system is presented. The proposed system can provide grid service policies for reactive power control according to the requirements specified in the IEEE 1547 standard. This model-free DRL approach can be adapted to other microgrid systems.**

I. Introduction

The rapid and deep penetration of distributed energy resources (DER) technologies, especially the electric vehicles (EVs) and the related inverter-based DER, has become more critical to the reliability and resilience of the microgrid. DC fast chargers have gained increasing popularity due to their high efficiency. The power controller within the charger is able to receive the grid service reference command to compensate for grid voltage/frequency fluctuations [1]–[3]. The IEEE Standard 1547 interconnects the DER with the grid by specifying the requirements relevant to performance, safety, and the maintenance of interconnection [4]. Recently, deep reinforcement learning (DRL) has been used as an advanced tool to solve complex optimal power flow and grid optimization problems due to its ability to make the decision in dynamic environments using an unsupervised approach [5], [6]. This paper proposes a DRL-based approach to optimize the microgrid and provide IEEE 1547-specified grid services implementation policies. The case study of optimizing the voltage fluctuation of a modified IEEE 30 bus system is demonstrated. The optimization results and evaluation show the effectiveness of the proposed DRL-based optimization approach with the Twin Delayed Deep Deterministic Policy Gradients (TD3) as agents to optimize the EV-interfaced microgrid under the IEEE 1547-specified grid service requirements. An online-updating architecture is demonstrated as the suggestion to use the proposed method in a real-world scenario. Our proposed approach can also be extended to optimize the performance of EV-interfaced microgrid, contribute to the Optimal Power Flow (OPF) problems, as well as manage the DC charger implementation strategies and EV dispatches.

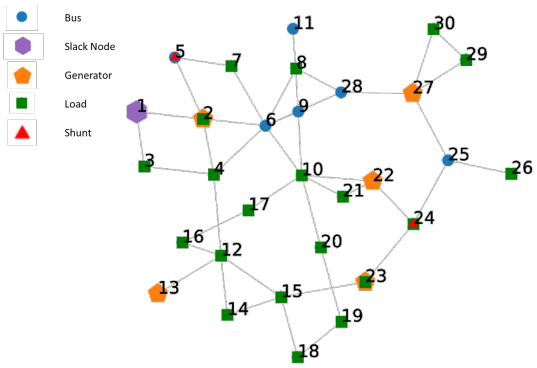


Figure 1: The network topology of the IEEE 30-bus system with all the load buses being capable of providing the IEEE-1547 specified grid services.

II. Method

A. Grid Structure

We consider an electric grid structure adapted from the IEEE 30-bus system, as shown in Figure 1, which consists of 30 buses, 1 slack node, 5 generators, 2 shunts, 42 transmission lines, and 20 loads. Each load represents a group of DER, including EVs and EV chargers, that could provide the IEEE 1547-specified grid services. To simulate the real-world hourly load changes on the 20 buses with load connected in the IEEE 30-bus system, 20 one-year load profiles from commercial, residential, and industrial buildings in Typical Meteorological Year 3 (TMY3) locations are used as the basis to modify the grid states [7]. The structure and bus information is obtained from *pandapower* [8].

B. IEEE 1547 Grid Services

The grid services following the IEEE 1547 standard mainly include six working modes [9]:

1) voltage-reactive power (Volt-Var) mode;
2) active-reactive power (Watt-Var) mode;
3) constant reactive power (Const-Var) mode;

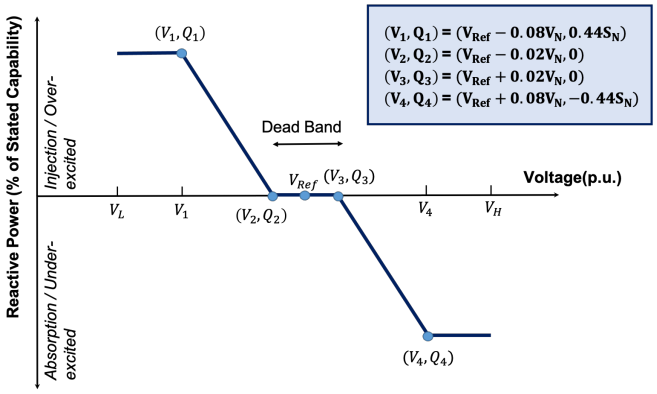


Figure 2: Volt-Var curve.

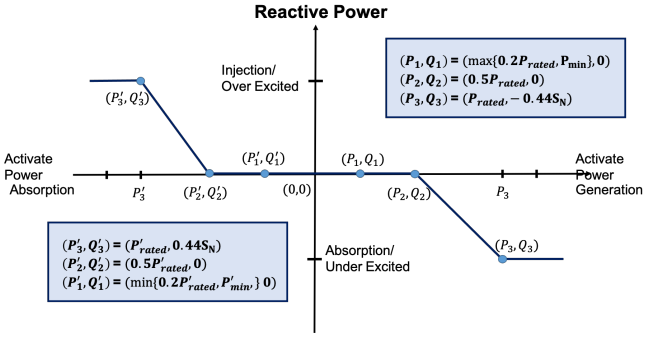


Figure 3: Watt-Var curve.

4) constant power factor (Const-PF) mode;
5) frequency-active power (Freq-Watt) mode;
6) voltage-active power (Volt-Watt) mode.

Considering the specifications in Annex B of IEEE 1547, the inverter-based DER is usually assigned as Category B DER, while synchronous machine generations belong to Category A DER [10]. This paper focuses on exploring the implementation policy of the reactive power control modes (*Const-Var*, *Volt-Var*, and *Watt-Var*) with Category B DER requirements since the chargers are able to conduct reactive power control with or without the EVs connected.

Specifically, the *Const-Var* mode targets at maintaining a constant injection or absorption of reactive power. The *Volt-Var* mode aims at adjusting reactive power to compensate for the voltage variation based on a piece-wise *Volt-Var* power response curve as shown in Figure 2. The definitions of the critical points are also denoted in the figure, where V_N is the nominal voltage, S_N is the nominal apparent power, and $\mathrm{V}_{\mathrm{Ref}}$ equals to the reference voltage (low pass filtered measured voltage). In the *Watt-Var* mode, the reactive power shall be governed by active power absorption or injection based on the predefined response curve as shown in Figure 3 in *Watt-Var* mode. P_{rated} and $P^{'}_{\mathrm{rated}}$ respectively denote the maximum active power can be injected and absorbed of the DER, while P_{min} and $P^{'}_{\mathrm{min}}$ respectively denote the minimum active power that can be injected and absorbed of the DER [11].

C. Deep Reinforcement Learning Based System Pipeline

1) Optimal Power Flow (OPF) Formulation: The environment to generate the rewards and define next step for the DRL is based on the AC power flow model. We use $G = (V, E)$ to denote the power grid topology, where $V = \{1, 2, \cdots, N\}$ represents the N buses and $E = \{e_1, e_2, \cdots, e_M\}$ represents the M transmission lines. The characteristics of a transmission line are determined by its admittance value: $y_{ij} = g_{ij} + \mathbf{j}b_{ij}$, where condutance and suspectance are denoted by $g_{ij} \in \mathbf{G}$ and $b_{ij} \in \mathbf{B}$, respectively. The active power, p_i, and reactive power, q_i, at bus i can be written as:

$$p_i = -\sum_{l \in V^i} |v_i| \cdot |v_l| \cdot (g_{il}\cos(\theta_i - \theta_l) + b_{il}\sin(\theta_i - \theta_l)), \quad (1)$$

$$q_i = -\sum_{l \in V^i} |v_i| \cdot |v_l| \cdot (g_{il}\sin(\theta_i - \theta_l) - b_{il}\cos(\theta_i - \theta_l)), \quad (2)$$

where V^i is the set of buses directly connected by edges to bus i, $|v_i|$ is the magnitude of the voltage at bus i, and θ_i is the phase angle at bus i.

The proposed DRL-based optimization approach for OPF problem can be utilized for different objectives, such as minimizing the generation cost or transmission loss. For the example case investigated in this work, the objective of DRL-based OPF is to stabilize the grid voltage variation on all buses while satisfying the grid operational constraints:

$$\min \sum_{i=1}^{N} |v_i| - |v_{ref,i}|, \quad (3)$$
$$\text{s.t., grid operational constraints.}$$

The objective function in Eq. (4) can be modified for other grid optimization purposes, such as minimizing the power loss on the transmission lines.

2) System Pipeline: Markov Decision Process (MDP) is usually used as the mathematical framework to describe an environment in the DRL problems. Our considered OPF problem can be modeled as an MDP with finite time steps containing four parts: $\langle S, A, P, R \rangle$. In particular, S represents the set of states, which is composed of the active power p_k and reactive power q_k at the bus k that connects with DER (loads). The action set $A = [\Delta q_1, \cdots, \Delta q_k]$ contains the incremental adjustment of reactive power injections (step-wise adjustment). P represents the transition probability to the next state, which is complex and strongly depends on the grid response modeled by the AC power flows, and the objective stated in Eq. (4). To address those issues, a model-free DRL-based approach is used to learn the transition procedure. R denotes the reward after an action is taken in a state.

Figure 4 shows the system pipeline to minimize the voltage variation on every bus in the microgrid, which uses the TD3 method as the RL agent. TD3 learns a deterministic policy

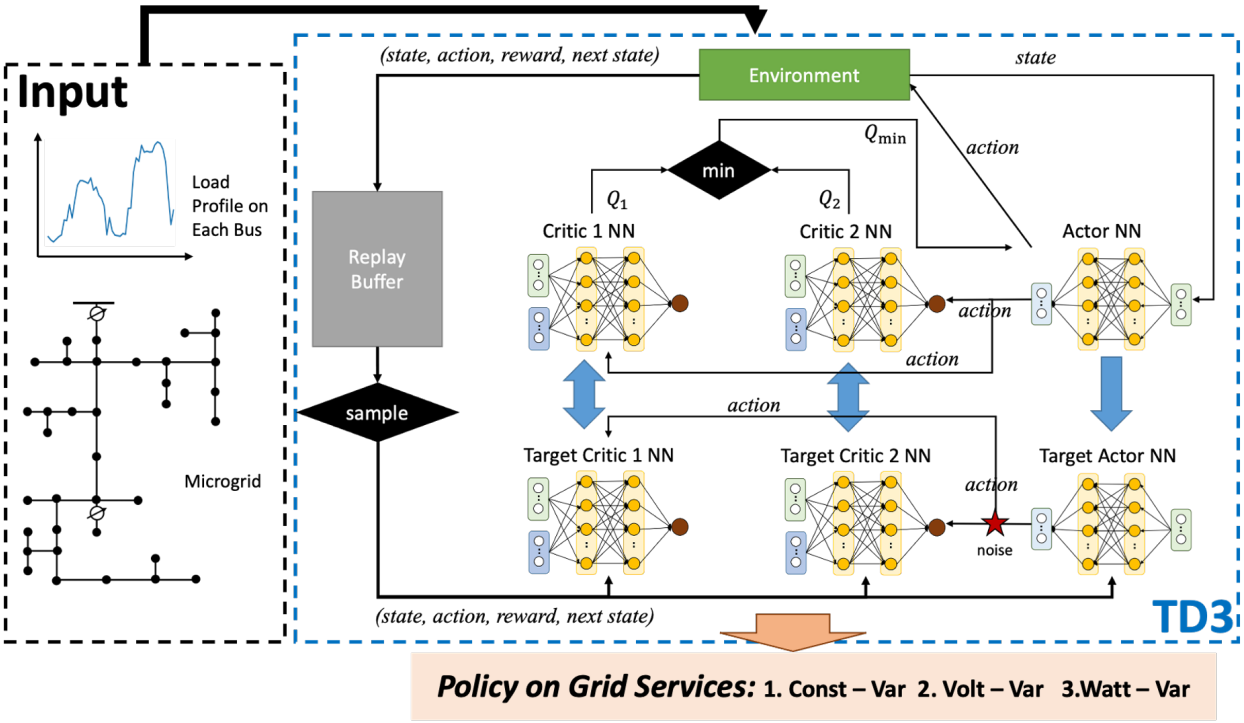


Figure 4: System pipeline for generating grid services policies by optimizing the microgrid using a DRL-based approach. (The blue arrows indicate the direction of parameter updates for the neural networks, and the red star represents the place to add noise during the training process to prevent overfitting.)

in an environment with continuous state and action spaces and is based on actor-critic algorithm [12]. The actor neural network (NN) is the policy function that maps state to action, while the critic NN maps the state to a scalar **Q** value that measures the quality of the input state. The characteristics of the grid, the OPF, and the load profiles are included in the environment, which is used to provide the grid information to calculate reward and derive the next state. The constraints of the three grid services modes mentioned before will be implemented in the environment in Figure 4.

Three DRL models are trained for the three reactive power grid service modes based on the IEEE 1547 requirements illustrated in Section II-B. The trained network can generate the optimal reactive power $q_{k,Optimal}$ on each bus that has DER connected. With $q_{k,Optimal}$, the resulting stabilized voltage $v_{i,optimal}$ can be derived. With that information, considering the real-world feasibility, the microgrid operator is able to decide which mode to follow and how much reactive power needs to be injected or absorbed at each bus with DER. The operator could also generate the grid service commands to the DER.

3) Training Reward and Grid Service Constraints: The limits of the voltage variation are $[v_{Lower}, v_{Upper}]$, which should be set by the grid operator according to the demand. For each epoch in every training episode, the environment based on the AC power flow equations is able to calculate the grid parameters (v_i, p_i, q_i, and θ_i) on every bus. The nominal reactive power of each DER is $q_{ref,k}$ A single-step reward R is empirically defined as:

$$D_i = \max\left(v_{Lower}^2 - v_i^2,\, v_i^2 - v_{Upper}^2,\, 0\right), \tag{4}$$

$$R = C - 10^{-8}\sum_{i=0}^{N}(q_k - q_{ref,k})^2 - 10^{-3}\sum_{i=0}^{K} D_i. \tag{5}$$

For each epoch in every training episode, the grid service constraints are added after executing the step-wise adjustment action. The constrained results will be used to calculate the reward. For each training episode, the episode ends when the single-step reward is larger than 0 or the calculated voltage for each bus falls within the predefined range $[v_{Lower}, v_{Upper}]$. The training process ends when the number of total steps in all episodes exceeds a certain amount (a value that depends on the complexity of the grid).

III. Results and Implementation

In this section, the performance of the proposed DRL-based method with TD3 is compared with other DRL agents, Deep Deterministic Policy Gradient (DDPG) and Deep Q Network (DQN). Specifically, DQN contains two Q-networks (a local NN and a target NN). While DQN is just a value-based learning method, DDPG is an actor-critic method with four NNs (2 actor NNs and critic NNs). In addition, TD3 learns two Q-functions and adds delays to update the policy to prevent the overestimation issues commonly happening with DDPG. As the agents, DDPG and DQN use the same environment

with continuous action spaces as the proposed method with TD3. In addition, an online-updating architecture is discussed to illustrate how the proposed method could be used in real-world scenarios.

A. Experimental Setup

Modified from the IEEE-30 bus case from *pandapower*, there are $N = 30$ buses and $K = 20$ DER. The nominal reactive power of DER are adjusted to represent a timestamp from the load profile where the voltage values on the 30 buses are not entirely scattered around the reference voltage $v_{ref,i} = 1, \forall i \in N$, as shown by the black dots in Figure 6. The lower and upper bounds are set to be $[v_{Lower}, v_{Upper}] = [0.975\,p.u., 1.25\,p.u.]$. The training process terminates when the total number of epochs in all episodes reaches $100,000$. The parameters used in the IEEE-1547 grid service V_N, S_N, P_{rated}, and P'_{rated} are set to be $1\,p.u.$, $43\,MVA$, $-40\,MW$, and $40\,MW$, respectively.

B. Training Results and Evaluation

The average reward curves for all epochs in each episode of *Constant-Var*, *Volt-Var*, and *Watt-Var* modes using the three algorithms are plotted in Figure 5. As shown in Figure 5, the agent cannot make good decisions at the beginning of the training process and needs to explore larger action spaces to achieve more information in each state. The agent finally learns the optimal policy by continuously interacting with the environment. The fluctuations in Figure 5 are caused by the noise added in the training process. Comparing the overall results in Figures 5(a), 5(b), and 5(c), TD3 and DDPG have significantly higher training rewards than DQN, while TD3 has slightly higher rewards than DDPG in the three reactive power grid services modes. In addition, comparing the training rewards for *Constant-Var* mode in Figures 5(a) and 5(b), DDPG has the problem being stuck at a local minimum.

Figure 6 illustrates the voltage before and after the three DRL optimization methods, where black, blue, red, and green dots represent the voltages without optimization, with *Watt-Var* mode, with *Volt-Var* mode, and with *Constant-Var* mode, respectively. Comparing Figures 6(a), 6(b), and 6(c), TD3 and DDPG perform better than DQN to learn the reactive power policy on each DER under different grid service modes to stabilize the voltage. Before implementing any DRL optimization, only 16 out of 30 buses are with the voltages in the predefined range. Table I shows the number of bus voltages that fall within $[0.975\,p.u.$ and $1.25\,p.u.]$ using the proposed method, DDPG method, and DQN method under three reactive power control modes specified by IEEE-1547. In addition, with the IEEE-30 bus topology and the example load information used in this paper, *Watt-Var* mode performs better in minimizing the fluctuations on every bus in the system. With different topology and load information from a different time stamp, another reactive power control mode may have the best performance for the optimization objective in Eq. (4).

Figure 7 presents the reactive power on each bus after adapting the reactive power policies under different modes and

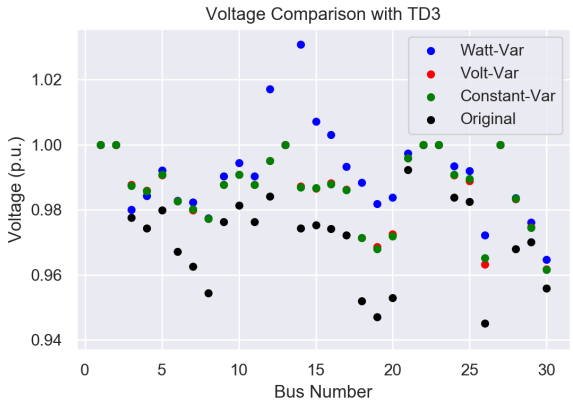
(a) Training Reward with TD3

(b) Training Reward with DDPG

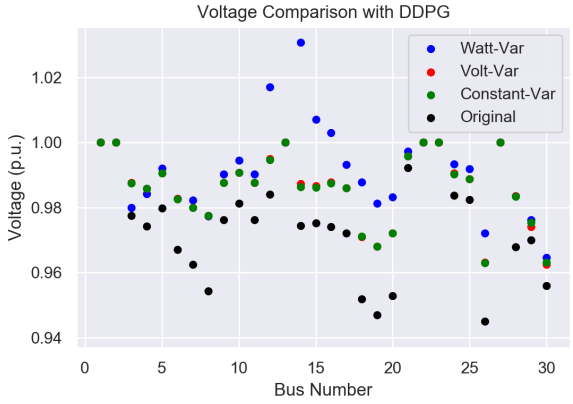
(c) Training Reward with DQN

Figure 5: The average training reward for all steps in each episode using TD3, DDPG, and DQN for *Constant-Var*, *Volt-Var*, and *Watt-Var* modes. (Green line: *Constant-Var*; Blue line: *Watt-Var*; Red line: *Volt-Var*.)

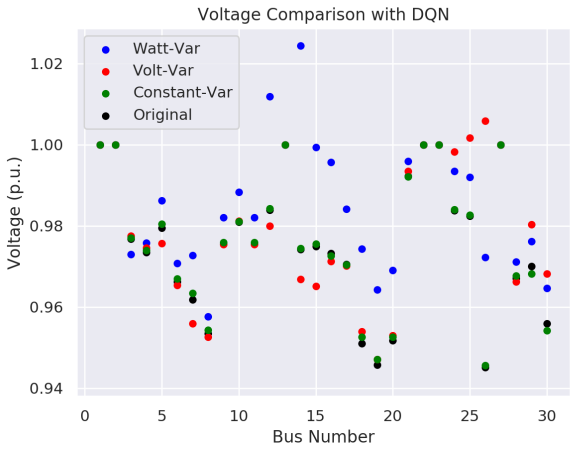
using different DRL optimization methods on the 20 buses that connect with DERs. In general, the proposed TD3-based DRL optimization method has the best performance among the three DRL approaches. DQN-based method performs poor



(a) Voltage Comparison with TD3



(b) Voltage Comparison with DDPG



(c) Voltage Comparison with DQN

Figure 6: The voltage optimization results on each bus for *Constant-Var*, *Volt-Var*, and *Watt-Var* modes using TD3, DDPG, and DQN. (Black dots: original voltage before optimization; Green dots: *Constant-Var*; Blue dots: *Watt-Var*; Red dots: *Volt-Var*.)

with continuous action spaces, and DDPG approach has the problem of overestimation on Q-value and overfitting.

Table I: Number of buses with voltage within the predefined range under three DRL methods and three grid service modes.

DRL Methods	Constant-Var	Volt-Var	Watt-Var
TD3	25	24	27
DDPG	24	24	27
DQN	16	17	20

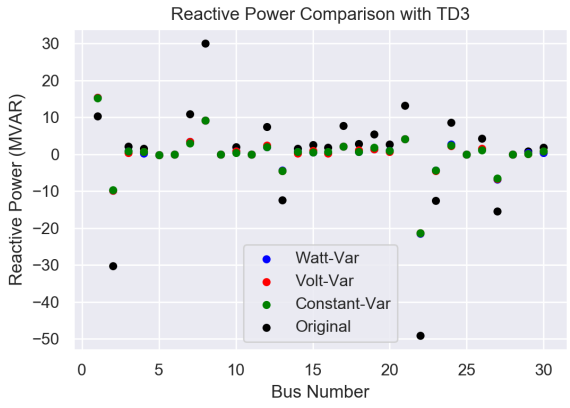
C. Online Update and Grid Services Implementation

As mentioned above, although the proposed DRL-based optimization method with TD3 as the agent performs well under different IEEE-1547 specified reactive power control modes, the specific mode that best optimizes the objective function highly depends on the current load information and grid topology. In addition, as the proposed DRL method has continuous action spaces, the TD3 agent will perform better after observing more conditions. Thus, we suggest an online-updating architecture to use the proposed method in the actual application scenarios shown in Figure 8.

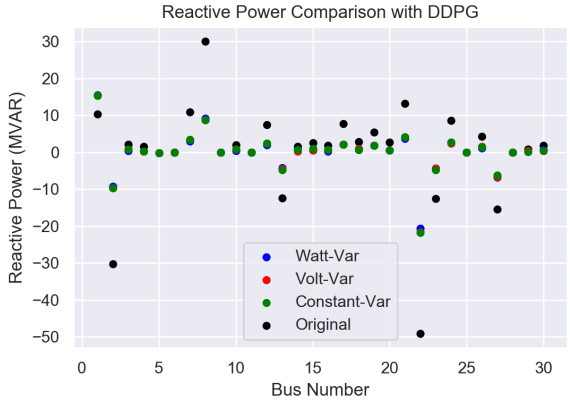
Specifically, historical data, which contains the load profile information from previous timestamps, can be used to generate the pre-trained models for *Constant-Var*, *Volt-Var*, and *Watt-Var* modes. The pre-trained models will serve as the operating models to provide optimization results and suggestions for grid service policies in a short time. When the grid operators need to come up with grid service commands with the current load profile, the current load profile can be directly used as the input for the pre-trained operating models as well as saved into the historical data. The historical data will be updated with new load profile information and utilized to train local grid services models. The local grid services DRL-based models using TD3 behave as the twins for the operating models with the same architecture. Every time when the local models finish training, the parameters can be passed to the operating models. With the online updating techniques, the operating models can keep improving their performance after meeting more load profiles. The output optimization results from the operating models can provide suggestions to the grid operators to generate the grid service commands to the DER. After the grid service commands are sent and implemented by the DER, the input load profile will change accordingly.

IV. Conclusion

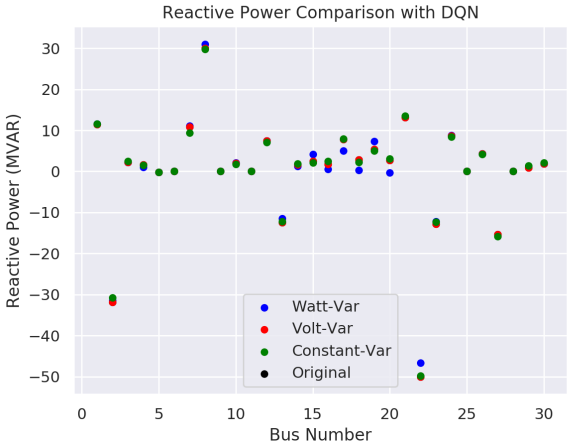
This paper proposes a method based on DRL with the TD3 algorithm to optimize the EV-interfaced microgrid considering the IEEE 1547-specified grid services requirements. The IEEE 30-bus is adopted as the microgrid structure in this work, where 20 out of 30 buses have EVs and EV chargers (DER) connected to the loads. The EV chargers are capable of implementing reactive control grid services. The voltage variation on each bus is regulated by implementing reactive power control policies generated by the DRL-based system. This model-free DRL-based approach is scalable to complex grid structures that might be challenging analyze using traditional approaches. It is also able to learn and generate the grid service policies for other objectives, such as minimizing the power loss in



(a) Reactive Power Comparison with TD3



(b) Reactive Power Comparison with DDPG



(c) Reactive Power Comparison with DQN

Figure 7: The resulting reactive power of each bus after being optimized by TD3, DDPG, and DQN for *Constant-Var*, *Volt-Var*, and *Watt-Var* modes. (Black dots: original voltage before optimization; Green dots: *Constant-Var*; Blue dots: *Watt-Var*; Red dots: *Volt-Var*.)

the microgrid. An online-updating architecture is shown to demonstrate how to efficiently use the proposed method in real-world scenarios to operate the grid. The proposed method

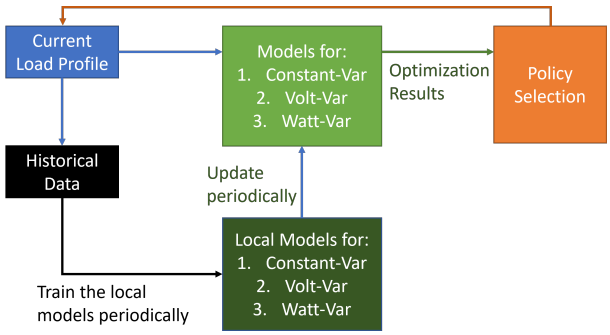


Figure 8: The architecture for online update for one grid structure in real application scenarios.

can optimize the performance of EV-interfaced microgrid and provide grid service commands suggestions to the grid operator. It also has the potential to offer insights into EV charging scheduling and EV charger implementation.

REFERENCES

[1] L. Zhou, M. Jahnes, M. Eull, W. Wang, G. Cen, and M. Preindl, "Robust control design for ride-through/trip of transformerless on-board bidirectional ev charger with variable-frequency critical-soft-switching," *IEEE Transactions on Industry Applications*, 2022.

[2] L. Zhou, M. Jahnes, M. Eull, W. Wang, and M. Preindl, "Control design of a 99% efficiency transformerless ev charger providing standardized grid services," *IEEE Transactions on Power Electronics*, 2021.

[3] J. Nie, L. Zhou, M. F. Kaye, *et al.*, "Optimal power flow estimation of microgrid considering the grid services of ev batteries," in *2021 IEEE Transportation Electrification Conference & Expo (ITEC)*, IEEE, 2021, pp. 249–254.

[4] *IEEE 1547-2018 - IEEE Standard for Interconnection and Interoperability of Distributed Energy Resources with Associated Electric Power Systems interfaces*, 2017.

[5] T. Sogabe, D. B. Malla, S. Takayama, *et al.*, "Smart grid optimization by deep reinforcement learning over discrete and continuous action space," in *2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC)(A Joint Conference of 45th IEEE PVSC, 28th PVSEC & 34th EU PVSEC)*, IEEE, 2018, pp. 3794–3796.

[6] D. Cao, W. Hu, X. Xu, *et al.*, "Deep reinforcement learning based approach for optimal power flow of distribution networks embedded with renewable energy and storage devices," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 5, pp. 1101–1110, 2021.

[7] E. Wilson, "Commercial and residential hourly load profiles for all tmy3 locations in the united states," DOE Open Energy Data Initiative (OEDI); National Renewable Energy Lab.(NREL . . ., Tech. Rep., 2014.

[8] L. Thurner, A. Scheidler, F. Schäfer, *et al.*, "Pandapower—an open-source python tool for convenient modeling, analysis, and optimization of electric power systems," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6510–6521, 2018.

[9] M. Ingram, R. Mahmud, and D. Narang, "Informative background on the interoperability requirements in ieee std 1547-2018," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2021.

[10] "Technical interconnection and interoperability requirements (TIIR)," Stearns Electric Association, Tech. Rep., 2021.

[11] H. V. Padullaparti, N. Ganta, and S. Santoso, "Voltage regulation at grid edge: Tuning of pv smart inverter control," in *2018 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*, IEEE, 2018, pp. 1–5.

[12] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International Conference on Machine Learning*, PMLR, 2018, pp. 1587–1596.