Soft Actor Critic Based Volt-VAR Co-optimization in Active Distribution Grids

Rakib Hossain †, *Student Member, IEEE*, Mukesh Gautam, *Student Member, IEEE*,
Mohammad MansourLakouraj, *Student Member, IEEE*, Hanif Livani, *Senior Member, IEEE*,
Mohammed Benidris, *Senior Member, IEEE*, and Yahia Baghzouz, *Senior Member, IEEE*
Department of Electrical & Biomedical Engineering, University of Nevada, Reno
Email: {rhossain†, mukesh.gautam, mansour}@nevada.unr.edu, {hlivani, mbenidris}@unr.edu, {yahia.baghzouz}@unlv.edu

***Abstract*—Modern distribution networks are undergoing several technical challenges, such as voltage fluctuations, because of high penetration of distributed energy resources (DERs). This paper proposes a deep reinforcement learning (DRL)-based Volt-VAR co-optimization technique for reducing voltage fluctuations as well as power loss under high penetration of DERs. In addition, the proposed approach minimizes the operational cost of the grid. A stochastic policy optimization based soft actor critic (SAC) agent is proposed to configure the optimal set-points of the reactive power outputs of the inverters. The performance of the proposed model is verified on the modified IEEE 34- and 123-bus systems and compared with a base case scenario with no reactive supply by inverters, and a local droop control approach. The results demonstrate that the proposed framework outperforms the conventional droop control method in improving the voltage profile, minimizing the network power loss, and reducing grid operational cost.**

***Index Terms*—Distribution grids, deep reinforcement learning, soft actor critic, Volt-VAR optimization.**

I. Introduction

With high penetration of distributed energy resources (DERs), modern distribution networks are experiencing different challenges such as large voltage fluctuations and increased power losses [1]. Moreover, uncertain generations from DERs and the wider range of generation fluctuations make it more challenging to operate the grid within the acceptable range of voltages. Volt/VAR optimization (VVO) is an efficient tool for system operators to address the challenges associated with the increasing penetration of intermittent resources. The main objective of VVO tools is to coordinate among different reactive power resources, such as smart inverters (SIs), and control their reactive power injection/absorption, to achieve a smooth voltage profile . Using the PV generation forecasts and according to the load profile, VVO can schedule the resources to keep the feeder's voltage profile within a permissible range under any possible loading and PV generation scenario. While maintaining the voltages in nominal level, the VVO framework can also minimize network power losses in the feeder [2].

In the literature, VVO techniques are generally categorized into three classes: a) numerical optimization, b) Heuristic optimization, and c) learning-based methods. The most widely used numerical optimization techniques are mixed-integer non-linear programming (MINLP) [3], mixed-integer linear programming (MILP) [4], and dynamic programming (DP) [5] that have been established based on optimal power flow analysis. These methods include various continuous variables and integer that might make them tedious for real-time execution. Some of the most widely used heuristic approaches for VVO are genetic algorithm (GA) [6], particle swarm optimization (PSO) [7], and teaching learning-algorithm (TLA) [8]. They have been employed explicitly for solving non-convex optimization problems with non-linear model of the distribution system. Although, these methods are fitted for complex scenarios, they can become impractical if number of decision variables increase. Moreover, whenever any new scenarios are encountered in the system, these optimization models are required to solve again, and also they cannot adopt properly with any drastic changes in the time-dependent loads and DERS in the network.

Learning-driven techniques have been implemented to address the limitations of aforementioned optimization approaches since they can handle uncertainties by extracting knowledge from historical data. Moreover, learning-driven models don't require to be solved whenever new scenarios is encountered because they can utilize their knowledge gained from previous experiences. Deep reinforcement learning (DRL) is one of the most suitable data-driven approaches with data exploration capabilities in non-linear high dimensional spaces using deep neural networks (DNNs) [9]. The authors in [10] applied a DRL technique to regulate voltage set-points of generators for controlling the voltage within the allowed range with a variety of line and load outages. The authors in [11] used deep deterministic policy gradient (DDPG) agent-based DRL approach to coordinate between different SIs to control the voltage profile of the grid. An attention enabled multi agent DRL (MADRL) is employed for the coordination among different PV inverters and static var compensators (SVCs) to control their optimal reactive power set-points to regulate the voltage profile of distribution system (DS) [12]. A consensus-based MADRL method is proposed to control the operational schedules of the utility devices like capacitor banks, and on-load tap changers (OLTCs) to regulate the voltage of DS [13]. The only objective of these DRL methods is to control the voltage profile of the network. This paper proposes a DRL-based framework for co-optimization of voltage regulation in the feeders and minimization of look-ahead operational cost of the distribution grid.

The main contributions of this paper are as below:

- It proposes a sample efficient DRL algorithm called soft actor critic (SAC) with continuous actions to learn a stochastic VVO policy. The proposed algorithm avoids potential instability and complexity associated with previous off-policy maximum entropy algorithm based soft Q learning. In high-dimension action space while off-policy based other DRL algorithms struggles, the proposed SAC algorithm can perform well.
- The proposed SAC agent coordinates among PV and battery energy storage (BES) inverters with their continuous reactive power outputs, and controls the active power charging/discharging the BESs based on the load demand.
- By optimal scheduling of intra-hour of smart inverter outputs, the proposed approach improves the voltage profile and reduces the power loss of the distribution system. Moreover, it reduces the operational cost of distribution grids as the second objective.
- The proposed VVO framework is formulated to adjust the inverters settings according to their allowed range of power factor changes, e.g., 0.9 leading/lagging. This constraint ensures that the efficiency of the inverters are kept within the specified limits of the manufacturers, and reduces the inverters' loss due to reactive power supply.

The rest of the paper is ordered as follows. Section II explains the framework. Section III justifies the performance of the model. Section IV provides concluding remarks.

II. The Proposed Volt-VAR Co-Optimization Framework

A. Soft Actor Critic (SAC) Algorithm

SAC is an off-policy reinforcement learning (RL) algorithm that optimizes a stochastic policy to maximize the long term entropy as well as expected lifetime rewards. It simultaneously learns a policy as well as two Q-functions, Q_{target_1}, Q_{target_2} , and applies the minimum of these two Q-values to build the targets in Bellman error functions as follows,

$$y(r, s, d) = r + \gamma(1-d)(min_{1,2}Q_{target,i}(s, a) - \alpha log\pi_\theta(a|s)) \tag{1}$$

where, d represents the done signal, α is a trade-off coefficient. and r denotes reward. In each state, the policy acts to maximize the expected future return together with expected future entropy i.e., it should optimize state value function, $V^\pi(s)$ which expand into

$$V^\pi(s) = E_{a\sim\pi}[Q^\pi(s, a) - \alpha log\pi(a|s)] \tag{2}$$

To optimize the policy, we have incorporated reparameterization trick, in which a sample is taken from the control actions as $a(t) \sim \pi_\theta(.|s)$ that is rendered by calculating a deterministic state function, independent noise, and policy parameters. We are incorporating squashed Gaussian policy, which indicates that samples obtained as

$$a_\theta(s, \epsilon) = tanh(\mu_\theta(s) + \sigma_\theta(s)\epsilon) \tag{3}$$

where $\epsilon \sim N(0, I)$, μ indicates the mean values of actions for a given state, σ represents the standard deviation. Although in deterministic policy based algorithms e.g., DDPG, and TD3, a random noise is added to next-state actions for smoothing of the target policy, in SAC based algorithm; no additional noise is required to add. Because SAC learns on stochastic policy and therefore the noise come from stochasticity is sufficient to obtain optimum control actions.

Algorithm 1: Proposed SAC Algorithm

```
Input: States from the environment (Voltages, $P_{loss}$)
Output: Actions to the environment ($Q_{pv}$, $Q_{BES}$, $P_{BES}$)
Initialize policy parameters, $\theta$ and Q-function parameters $\Phi_1$ and $\Phi_2$
Initialize replay buffer $D$
Set the target networks parameters equals primary parameters as $\Phi_{target,1} \leftarrow \Phi_1$, $\Phi'_{target,2} \leftarrow \Phi_2$ and
for episodes $1, 2, 3, \cdots, N$ do
  Initialize the power flow and take the initial states, $S_0$ from the environment
  for time-slot $t = 1, 2, 3, 5, \cdots, T$ do
    Derive the control actions using $a(t) \sim \pi_\theta(.|s)$
    Execute the control actions to the environment
    Get the rewards using (4)
    Update the state,$S'$, store the transition in replay memory, $D$.
    Randomly sample a batch of N transitions from, $D$ calculate the target Q functions using (1)
    Using gradient descent upate the Q function as $\Delta_{\theta_i}\frac{1}{|B|}\sum_{(s,a,r,s',d)\epsilon B}(Q_{\Phi_{target_i}}(s, a) - y(r, s', d))^2$ for $i = 1, 2$ the gradient ascent update the policy as $\Delta_\theta\frac{1}{|B|}\sum_{s\epsilon B}(Q_{\Phi_i}(s, a_\theta(s)) - \alpha\log\pi_\theta(a_\theta(s)|s) - y(r, s', d))^2$ for $i = 1, 2$
    Update the target network as $\Phi_{target_i} \leftarrow \beta\Phi_{target_i} + (1-\beta)\Phi_i$ for $i = 1, 2$
  end for
  if no voltage violations, or reward converges and reached to the maximum iteration then
    BREAK
  end if
end for
```

B. Implementation of SAC

The SAC agent coordinates among DERs to provide fast and effective actions. The agent gets its reward based on the actions taken for a particular state of the environment. The details of the learning process is shown in Algorithm 1. The definitions of states, actions, and rewards are listed below:

- The states, s, are defined as a vector of measurements that represents environment conditions. In the proposed VVO model, we have considered voltage of each nodes and power loss of the network as state input.
- Based on the system state, the agent takes an action. The action space represents the control actions taken by an agent that is structured with several control variables. The proposed method uses reactive powers of PV inverters

the agent.

and BESs as well as active powers of BESs as actions of the agent.
- The reward function is a critical part for the evaluation of action-value, representing the effects of control actions to the environment states. We have considered both power loss, voltage fluctuation, and operational cost objectives in our reward functions which is shown in (4)

$$R(t) = -\eta_P \times (P_{loss}(t) - P_0) + \eta_v - \eta_{op} \times C_{op} \quad (4)$$

where $R(t)$ is the reward at time t; $P_{loss}(t)$ denotes the network power loss at time t based on the actions taken by the agents; P_0 represents the loss for any default action taken at time t_0. η_p , η_v , and η_{op} are the incentive factors for mitigating power loss, voltage fluctuations, and operational cost respectively. The value of η_p and η_{op} are chosen as 3 and 1 respectively while Table I shows variation of η_v under different voltage conditions. C_{op} indicates the operational cost that is calculated by (5).

TABLE I
INCENTIVE FACTORS AT DIFFERENT VOLTAGES

Conditions No	$Vmax$	$Vmin$	η_v
1	< 1.05	> 0.95	$+20$
2	< 1.05	> 0.9 but < 0.95	-5
3	< 1.05	< 0.9	-10
4	> 1.05 but < 1.1	< 0.95	-5
5	> 1.1	< 0.95	-10
6	> 1.05	< 0.95	-20

III. Case Study

The performance of the proposed intelligent agent is validated on the modified IEEE 34- and IEEE 123-node test feeders. The load and PV profiles are taken from a real-world dataset with a PV installation site at Henderson, Nevada, USA [14]. The load and PV profiles for both IEEE 34 and 123-bus systems are shown in Fig. 1 and Fig. 2 respectively. In the IEEE 34-bus test system, nine aggregated PV inverters with a total maximum generation of 52% of the total demand and four aggregated BESs inverters with a total maximum capacity of 37% of the total PV generation are installed on the primary feeder. In the 123-bus test case, seven aggregated PV inverters with a maximum generation of 42% of the total demand and four aggrgeated BESs inverters with a total capacity of 41% of the total PV generation are installed on the feeder. The details about the sizes and locations of both distribution systems are summarized in Table III, and Table IV. The inverter reactive powers are controlled so that they can operate at a power factor greater than or equal to 0.9 p.u.

A. Agent Training and Learning Process

To learn the optimal policy for delivering the best control action, the agent has been trained for 500 episodes. The actor-network learns the policy to get the optimal actions while the critic network learns the Q-values. These two networks are the fully connected neural networks (FCNN) that include the input layer, hidden layers, and output layer. Table V shows details about the DNN hyperparameters. At the starting phases of the training process, the agent randomly explores the decision space of the environment, and it eventually converges and obtains the optimal actions to minimize voltage fluctuations and power losses of the network. The cumulative reward in each episode is enumerated by summing the rewards from each time slot before the training process progress to the next episode. Initially, the rewards are low because, during these phases, the agent doesn't have previous knowledge about how to regulate the voltage. As the learning continues, the agent learns from the previous experiences. Therefore, the rewards increase and the agent converges after a certain number of episodes and gets the maximum rewards as shown in Fig. 3.

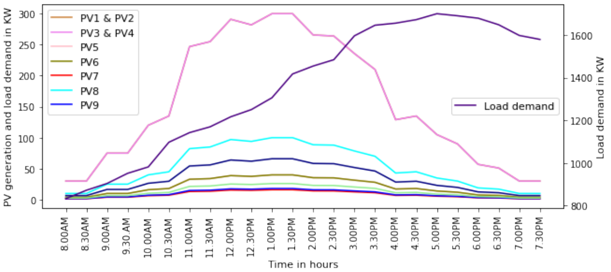

Fig. 1. The PV and load profiles of 34 bus system

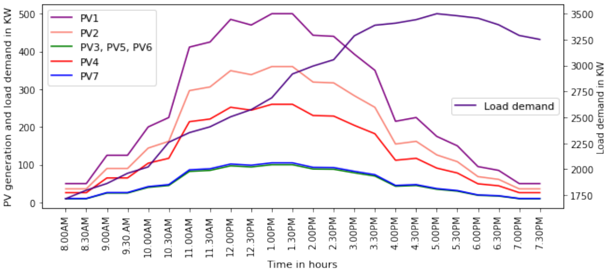

Fig. 2. The PV and load profiles for 123 bus system

TABLE II
MODIFIED IEEE TEST FEEDERS PARAMETERS

Parameters	34-node test case	123-node test case
No of PVs	9	7
No of batteries	4	4
PV penetrations	52% of peak load	42% of peak load
Battery penetrations	37% of PV rating	41% of PV rating
Maximum demand	1700 kW	3650 kW

TABLE III
SIZE AND LOCATION OF SIS FOR THE MODIFIED 34-BUS TEST SYSTEM

DERs	Location	Max active power(KW)	Phase
$PV1$	890	300	3
$PV2$	844	300	3
$PV3$	860	40	3
$PV4$	848	40	3
$PV5$	830	16	1
$PV6$	mid 822.1	100	1
$PV7$	mid 806.2	18	1
$PV8$	mid 836.3.1	26	1
$PV9$	mid 860.3.1	66	1
$BES1$	890	125	3
$BES2$	mid 840.2.3	10	1
$BES3$	844	125	3
$BES4$	mid 836.1.2	50	1

B. VVO Performance

1) Voltage Regulation (VR): One of the objectives of the proposed approach is to minimize voltage fluctuation prob-

TABLE IV

SIZE AND LOCATION OF SIS FOR THE MODIFIED 123-BUS TEST SYSTEM

DERs	Location	Max active power(KW)	Phase
$PV1$	48	500	3
$PV2$	47	360	3
$PV3$	76.1.2	100	1
$PV4$	76.1.3	260	1
$PV5$	65.3.1	100	1
$PV6$	2.2	100	1
$PV7$	7.1	105	1
$BES1$	48	250	3
$BES2$	47	150	3
$BES3$	76.1.2	100	1
$BES4$	35.1.2	130	1

TABLE V

DRL PARAMETERS SETTING

Parameters	Values
Hidden layers (actor network)	2
Hidden layers (critic network)	2
Activation function (output layer)	$tanh$
Activation function (hidden layers)	$ReLU$
No. of neurons in hidden layers	25, 25
Learning rate (actor)	10^{-3}
Learning rate (critic)	10^{-3}
Discount factor	0.99
Optimizer	Adam

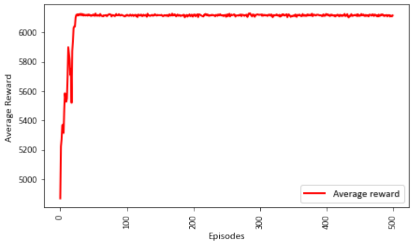

Fig. 3. Average scores in the training process

lems in distribution grids. The proposed DRL agent coordinates among the BES and PV inverters for optimal control of reactive power set-points and the active power charging/discharging of the BESs. During the morning hours till early afternoon, the BESs are charged to avoid the over-voltage since the PV generation usually exceeds the demand during this period. Beginning from early afternoon till the end of the scheduling period, the batteries are scheduled to discharge their energy to meet the overload in order to avoid under-voltage scenarios. Fig. 4 and Fig. 6 illustrate the minimum voltages of the network over the entire scheduling period for both IEEE 34-bus and 123-bus test cases respectively. Fig. 4 shows that in the base case, where the agent doesn't take any action, the minimum voltage of the network falls below the 0.93 p.u level. According to ANSI limits , this is a violation of standard voltage level [0.95-1.05 p.u.]. However, the proposed DRL agent can regulate the minimum voltage across the network close to the nominal value of 1.0 p.u. Fig. 5 depicts the voltage variations in each node at time 2.00 PM using different scheduling and control methods. It can be observed, in the base case scenario, the voltage fluctuation is high as there is no reactive power control in the system. Although local droop control method can adjust the voltages over a limited range, their capability in improving the voltage profile along the feeder is not completely utilized due to lack of sufficient coordination. The proposed SAC agent-based approach demonstrates better performance in regulating the feeder voltages compared to other approaches as it can effectively coordinate among the participating inverters for reactive power supply. The scalability of the proposed method is validated on a modified IEEE 123-bus test case. In this test case also, the proposed VVO framework outperforms other approaches in respect of improving the voltage profile as shown in Fig. 6.

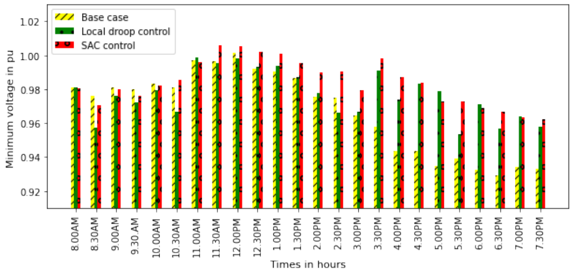

Fig. 4. Minimum voltage for the modified IEEE 34-bus test case

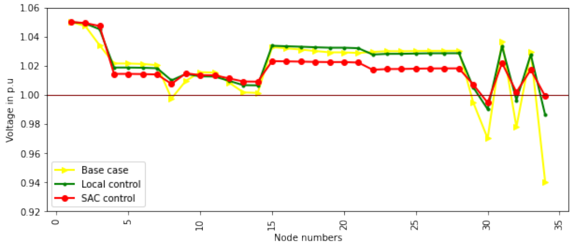

Fig. 5. Node voltages the modified IEEE 34-bus system at 2.00PM

2) Power Loss Minimization: The proposed agent is also designed to reduce power losses simultaneously with voltage fluctuation minimization by controlling active and reactive power of BESs and reactive power of PV inverters. Fig. 7 and Fig. 8 represent the power losses for both IEEE 34- and IEEE 123-bus test cases for all three scenarios. These two figures demonstrate that compared to local droop control, DRL agent based VVO approach has less power loss that justifies the performance of the proposed method. This is because the proposed DRL agent can effectively control the output power injection/absorption by the DER inverters.

C. Operational Cost Minimization

Minimization of operational cost is another objective in our co-optimization scheme that is formulated using (5),

$$OC(t) = \sum_{t=1}^{24} \sum_{i=1}^{N_{dg}} \alpha_{dg_i} P_{dg_i}^{t} + \sum_{t=1}^{24} \sum_{n=1}^{N_{sub}} C_{sub_n}^{t} P_{sub_n}^{t} \qquad (5)$$

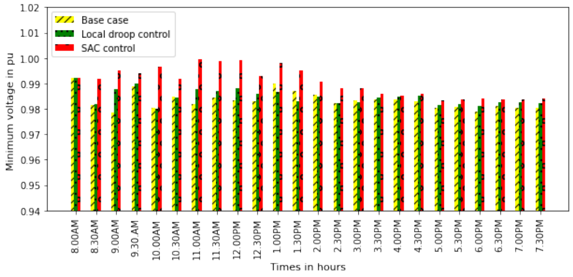

Fig. 6. Minimum voltage for the modified IEEE 123-bus test case

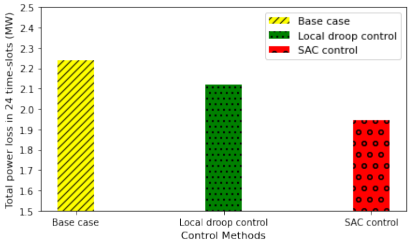

Fig. 7. Total power loss for the modified IEEE 34-bus test case

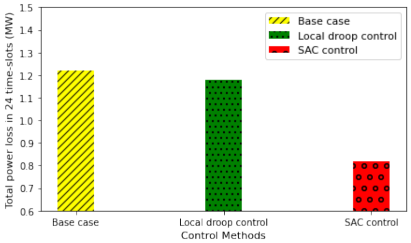

Fig. 8. Total power loss for the modified IEEE 123-bus test case

where $P^t_{dg_i}$ is real power of the i_{th} DG and $P^t_{sub_n}$ denotes active power of the n_{th} sub-station at time, t. The total operational cost for the IEEE 34-bus and 123-bus test cases during the entire scheduling period is shown in Fig. 9 and Fig. 10 respectively. We have compared the performance of the proposed DRL agent with the base case and droop control method. The figures demonstrate that the proposed framework converges at the lowest operational cost for the network compared to the other approaches, which verifies its ability of co-optimizing different grid services along with VVO.

IV. Conclusion

This paper has proposed an off-policy maximum entropy, actor-critic method called soft-actor critic (SAC) for VVO co-optimization in distribution grid with inverter-based resources. The agent interacts with the environment and adaptively chooses the optimal active/reactive schedules of BESs inverters and reactive power schedules of PV inverters to regulate grid voltages and reduce power losses in the network. Moreover, the framework minimizes the grid operational cost and the added inverter power losses. The inverters reactive powers were limited by the maximum range of allowed power factor in lagging/leading mode, i.e., 0.9. In high-dimensional action space, while other off-policy-based algorithms such as DDPG, are highly sensitive to hyperparameters and need more tuning for the convergence, the proposed SAC agent was less sensitive to the choice of hyperparameters with higher convergence rate. The performance of the proposed framework was compared with the base case scenario without any VVO, and with a local droop control of the inverters using two modified distribution feeders, the IEEE 34- and 123-bus systems. The simulation results validated the superior performance of the proposed method compared to the other optimization approaches, in terms of improving the voltage profiles, reducing the network power losses and minimizing the operational cost.

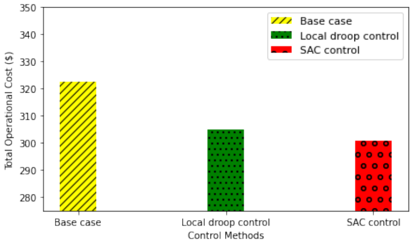

Fig. 9. Operational cost for the modified IEEE 34-bus test case

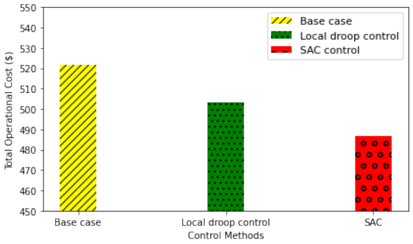

Fig. 10. Operational cost for the modified IEEE 123-bus test case

Acknowledgement

This paper is based upon work supported by the U.S. DoE's Office of EERE under the Solar Energy Technologies Office Award Number DE-EE0009022.

References

[1] T. Ding, C. Li, Y. Yang, J. Jiang, Z. Bie, and F. Blaabjerg, "A two-stage robust optimization for centralized-optimal dispatch of photovoltaic inverters in active distribution networks," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 2, pp. 744–754, 2017.

[2] V. Sarfi and H. Livani, "Optimal Volt/VAR control in distribution systems with prosumer DERs," *Electric Power Systems Research*, vol. 188, p. 106520, 2020.

[3] M. S. Hossan, B. Chowdhury, M. Arora, and C. Lim, "Effective CVR planning with smart DGs using MINLP," in *2017 NAPS*. IEEE, 2017.

[4] A. Borghetti, "Using mixed integer programming for the Volt/Var optimization in distribution feeders," *Electric Power Systems Research*, vol. 98, pp. 39–50, 2013.

[5] F.-C. Lu and Y.-Y. Hsu, "Reactive power/voltage control in a distribution substation using dynamic programming," *IEE Proceedings-Generation, Transmission and Distribution*, vol. 142, no. 6, pp. 639–645, 1995.

[6] S. Liu, J. Zhang, Z. Liu, and H. Wang, "Reactive power optimization and voltage control using an improved genetic algorithm," in *2010 International Conference on Power System Technology*. IEEE, 2010.

[7] H. Yoshida, K. Kawata, Y. Fukuyama, S. Takayama, and Y. Nakanishi, "A particle swarm optimization for reactive power and voltage control considering voltage security assessment," *IEEE Transactions on power systems*, vol. 15, no. 4, pp. 1232–1239, 2000.

[8] T. Niknam, M. Zare, and J. Aghaei, "Scenario-based multiobjective Volt/Var control in distribution networks including renewable energy sources," *IEEE Transactions on Power Delivery*, vol. 27, no. 4, pp. 2004–2019, 2012.

[9] 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:2006.00546*, 2020.

[10] 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 PESGM*. IEEE, 2019, pp. 1–5.

[11] C. Li, C. Jin, and R. Sharma, "Coordination of PV smart inverters using deep reinforcement learning for grid voltage regulation," in *2019 18th ICMLA*. IEEE, 2019, pp. 1930–1937.

[12] D. Cao, J. Zhao, W. Hu, F. Ding, Q. Huang, and Z. Chen, "Attention enabled multi-agent DRL for decentralized Volt-VAR control of active distribution system using PV inverters and SVCs," *IEEE Transactions on Sustainable Energy*, 2021.

[13] Y. Gao, W. Wang, and N. Yu, "Consensus multi-agent reinforcement learning for Volt-Var control in power distribution networks," *IEEE Transactions on Smart Grid*, 2021.

[14] "Solar Forecast Arbiter," accessed: 2021-08-30. [Online]. Available: https://dashboard.solarforecastarbiter.org/