A Multi-Agent Deep Reinforcement Learning Based Voltage Regulation Using Coordinated PV Inverters

Cao, Di; Hu, Weihao; Zhao, Junbo; Huang, Qi; Chen, Zhe; Blaabjerg, Frede

Published in:
I E E E Transactions on Power Electronics

DOI (link to publication from Publisher):
10.1109/TPWRS.2020.3000652

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2020.3000652, IEEE Transactions on Power Systems

A Multi-agent Deep Reinforcement Learning Based Voltage Regulation using Coordinated PV Inverters

Di Cao, *Student Member, IEEE*, Weihao Hu, *Senior Member, IEEE,* Junbo Zhao, *Senior Member, IEEE*, Qi Huang, *Senior Member, IEEE,* Zhe Chen, *Fellow, IEEE*, Frede Blaabjerg, *Fellow, IEEE*

***Abstract*—This paper proposes a multi-agent deep reinforcement learning-based approach for distribution system voltage regulation with high penetration of photovoltaics (PVs). The designed agents can learn the coordinated control strategies from historical data through the counter-training of local policy networks and centric critic networks. The learned strategies allow us to perform online coordinated control. Comparative results with other methods show the enhanced control capability of the proposed method under various conditions.**

***Index Terms*—Voltage regulation, multi-agent deep reinforcement learning, coordinated control, distribution system.**

I. Introduction

The increasing penetration of photovoltaics (PVs) in the distribution network (DN) may cause swing of voltages due to their rapid power variations. To solve this problem, various approaches have been developed. For the control strategies, the voltage control strategies can be divided into active power-based and reactive power-based methods. Active power-based methods are suitable for voltage regulation of low voltage DN with high R/X ratio. Charging scheduling of battery storage systems [1] and power curtailment of PVs [2] are two main strategies. Active power curtailment of PV generations reduces the absorption capacity of DN for solar energy. By contrast, the charging/discharging control of battery energy systems is costly. On the other hand, reactive power-based strategies reduce the fluctuation of voltage by utilizing the voltage control ability of capacitor banks, static var compensator and PV inverters. As shown in [3-4], reactive power control is an effective and economic way for the voltage regulation. From the perspective of control frameworks, the voltage regulation strategies can be classified into three categories [5]: centralized [6], local control without communications [7] and decentralized implementation with communications [8-9]. The centralized control strategies require fast and reliable communication links, which is challenging for practical distribution systems. The local control approaches can make decisions based on local observations, but the capability of resources may not be fully utilized due to the lack of cooperation between agents. The decentralized and coordinated control methods can achieve cooperation control using local information with limited communication links [9].

In recent years, with the development of artificial intelligence, the multi-agent deep reinforcement learning (MADRL) algorithm is becoming popular for various applications. In the MADRL algorithm, control units are modeled as intelligent agents with different control strategies. The agents can develop the optimal control strategies through the interactions with the environment and learn co-operation by modeling policies of other agents during offline training. When the training is done, the agents can provide decisions that have strong adaptability for the unknown dynamic in real-time.

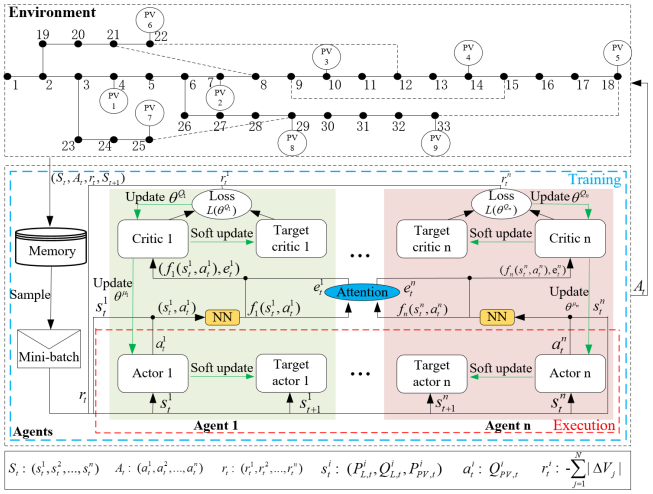

Fig. 1. The architecture of the proposed coordinated PV control approach.

This paper proposes to extend the MADRL based approach for the coordinated control of PV inverters. It has several benefits: 1) only local information is needed and the deployment of costly communication devices is not required. This distinguishes the widely used machine learning-based centralized control framework that is computational expensive; 2) the proposed MADRL based method is able to extract an optimal coordinated control strategy from historical data. This information embedded with past operational experience can generalize to newly encountered situations without resolving the optimization problem. The decision procedure is similar to recalling the past control experience from memory, which allows it to be implemented online; 3) Compared with [10], the attention model is integrated with MADDPG algorithm in this paper to enhance its scalability for more control subjects; Furthermore, this paper addresses the distribution system voltage regulation with high penetration of PVs instead of the transmission system voltage control.

II. Proposed Control Methodology

In this section, the voltage regulation problem in DN is first formulated as Markov games. This is solved by the proposed MADDPG algorithm with attention model.

A. Formulation of Markov Games

The architecture of the proposed MADRL based approach for voltage regulation is shown in Fig. 1. In the proposed control framework, the DN is the environment and each PV inverter serves as an agent. The cooperative control of PV inverters can be modeled as Markov games for N agents, including

• *The state set*: $s_t^i \in S_t$ is the local information obtained by agent i at time step t. S_t denotes the state of all agents at time step t. In this context, s_t^i consists of $P_{L,t}^i$, $Q_{L,t}^i$ and $P_{PV,t}^i$, which represent the active and reactive power of load demand of the node that the agent is connected to, and the active power injection of PV, respectively.

• *The action set*: $a_t^i \in A_t$ denotes the action of agent i at time step t. A_t represents the actions of all agents at time step t. In this context, the action is $Q_{PV,t}^i$, which represents the reactive power of the corresponding PV inverter.

• *The reward function*: $r_t^i \in r_t$ denotes the immediate reward for agent i at time step t. r_t is the set of rewards for all agents. In this paper, the reward is defined as: $r_t^i = -\sum_{j=1}^{N} |\Delta V_j|$, where ΔV_j and N are the voltage deviation of bus j and the number of buses, respectively.

At each time-step, every agent makes an action a_t^i based on the local observation s_t^i and obtains an immediate reward r_t^i, which relies on the global state and actions of other agents. When all the agents complete the actions, the system transfers to the next state, i.e., the Markov games. The objective of each agent is to learn a coordinated policy that maps s_t^i to a_t^i so as to maximize the obtained reward.

B. MADDPG with Attention Model for Markov Games

To achieve the coordinated policy, an actor-critic based MADRL approach is proposed. Specifically, each agent is composed of an actor μ_i and a critic Q_i. The actor that maps s_t^i to a_t^i is the policy function. The critic that maps (S_t, A_t) to a scalar is the judgment of the action made by this actor based on the global state. The actor and critic of each agent are trained against each other such that the critic can provide better evaluation and the actor can produce reactive power with reduced voltage deviation.

1) Actor:

In this paper, the neural network (NN) parameterized by θ^{μ_i} is used to approximate the policy function to deal with the system dynamics. The parameters θ^{μ_i} is optimized according to the gradient of the following performance function [11]:

$$\nabla_{\theta^{\mu_i}} J(\mu_i) = E_{S_t, A_t \sim D}[\nabla_{\theta^{\mu_i}} \mu_i(a_t^i \mid s_t^i) \nabla_{a_t^i} Q_i^{\mu}(S_t, a_t^1, ..., a_t^N)|_{a_t^i = \mu_i(s_t^i)}] \quad (1)$$

2) Critic with Attention Model

In the MADDPG algorithm, the states and actions of all agents are treated equally by the critic. This leads to two disadvantages: the spatial properties between different agents are ignored and the input of the critic grows linearly with the agent number, which may cause a performance degradation when applied to a large system with a large number of control objects. To address that, the attention critic is developed. It allows intelligently learning to attend to specific information that is most relevant to the rewards. This mechanism enhances the scalability of original MADDPG algorithm to deal with scenarios with more control objects. $Q_i^{\mu_i}(S_t, A_t)$, the critic of agent i, is a function of all agents' states and actions, and is expressed as follows [12]:

$$Q_i^{\mu_i}(S_t, A_t) = g_i(f_i(s_t^i, a_t^i), e_t^i) \quad (2)$$

where $f_i(\cdot)$ is the embedding function of agent i composed of a one-layer NN; $g_i(\cdot)$ is a two-layer NN used to approximate the critic function; e_t^i is the weighted sum of the contribution of all agents except for the agent i:

$$e_t^i = \sum_{j \neq i} \alpha_t^j \cdot u_t^j \quad (3)$$

$$u_t^j = V^T \times f_j(s_t^j, a_t^j) \quad (4)$$

$$\alpha_t^j \propto \exp((u_t^j)^T W_k^T W_q f_i(s_t^i, a_t^i)) \quad (5)$$

where α_t^j is the attention weight that agent i pays for agent j at time step t; u_t^j is the embedding of j's state and action; V^T is a matrix used for linear transformation; $f_j(\cdot)$ is the embedding function of agent j; the attention weight α_t^j is derived by comparing the embedding value of agent i and agent j; W_k and W_t are transition matrices. The parameters of the attention model (parameters in equation (3-5)) and of the critic function for agent i are represented by θ^{Q_i}. They will be optimized in a supervised fashion [11].

3) Target Networks and Replay Buffer

Target actor networks parameterized by $\theta^{\mu_i'}$ and target critic networks parameterized by $\theta^{Q_i'}$ are introduced to stabilize the training process. The replay buffer mechanism is also used to break the correlation between the data [8]. The parameters of NN of agent i can be denoted as $\{\theta^{\mu_i}, \theta^{Q_i}, \theta^{\mu_i'}, \theta^{Q_i'}\}$.

4) Implementation of the Proposed Method

The implementation of the proposed approach can be divided into two stages: centralized offline training and decentralized online execution. In the training stage, each agent is composed of actor networks and critic networks. The actor networks take local information s_t^i as input while the input of critic networks is augmented with state and action of other agents. This helps the agent model the decision procedure of other agents and contributes to the formulation of a cooperative control strategy. Because the training is done in off-line simulations, the information exchange can be achieved without specific communications. During the training process, the actor and the critic are trained against each other to learn an optimal control strategy. For detailed parameter optimization procedure, please refer to [11]. When the training is completed, the parameters of networks are fixed and only the actor is kept. Then, the actor of each agent can make decisions in real-time based on local observations. Since the augmented information is only used by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2020.3000652, IEEE Transactions on Power Systems

TABLE I Real-time decentralized control algorithm

Algorithm Decentralized real-time reactive power scheduling

1: Load the parameters of actor network of each agent θ^{μ_i}
2: for time step t=1,2,…T do
3: for agent i = 1,…, N do
4: obtain the local observation s_t^i
5: calculate action a_t^i according to $a_t^i = \mu_i(s_t^i \mid \theta^{\mu_i})$
6: output action of agent i a_t^i
9: end for
10: concatenate actions of all agents $A_t = (a_t^1, ..., a_t^N)$
11: end for
12: **Return**: $A_1 : A_T$

the critic during training, the agents can exhibit cooperative behaviors and provide decisions that are robust to other agents' decisions using local information. The procedures of the real-time decentralized control algorithm are shown in Table I.

III. Case Study

A. Simulation Setup

Simulations are carried out on the IEEE 33-bus system to verify the effectiveness of the proposed method. The system configuration is shown in Fig. 1. One-year PV output data from Xiaojin, a county in the Sichuan province of China is scaled and used for validation. The data are divided into training and test sets. The training set is used to train the agents and learn a cooperative control strategy based on local observations while the test set is applied to investigate the performance of the mastered control strategies. The simulation setup is shown in Table II. The number of agents is set as 9, each corresponding to a PV inverter. Every agent has two actor networks and two critic networks. All the networks share the same structure. The numbers of hidden layers are 100 and 100, respectively. The parameters setting of the proposed method are shown in Table III. Several methods from the literature are compared to demonstrate the achieved benefits from the proposed method, including the traditional droop control method and the

TABLE II Simulation setup

Maximum voltage deviation	5%
Rated power of PV (MW)	1.5
Apparent power of PV (MVA)	1.575

Table III Parameter settings of the proposed method

Parameters	Values
Batch size for updating NN	32
Replay buffer size	48000
Discount factor	0
Soft update coefficient	0.001
Learning rate for actor network	0.001
Learning rate for critic network	0.002

TABLE IV Voltage deviation of various methods on test data

Method	Average	Max rise	Max drop	Calculation time (s)
Original	1.68%	7.42%	3.66%	-
Droop	0.46%	3.48%	1.47%	0.0002
MADDPG	0.16%	1.19%	0.87%	0.0012
Proposed	0.12%	1.01%	0.75%	0.0012
Centralized	0.04%	0.62%	0.60%	0.67

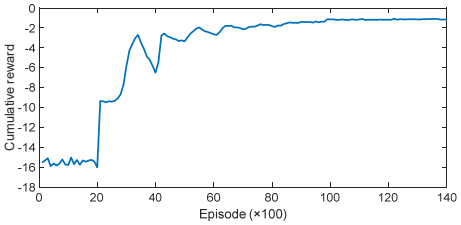

Fig. 2. The change of the reward during the training procedure

MADDPG-based method. It is worth noting that it is the first time the MADDPG-based method is extended for distribution system voltage control using PV inverters. Our proposed method further improves that with the attention strategy.

B. Training Process

The proposed approach is trained for 14000 episodes to learn the optimal voltage regulation strategy. An episode includes 24 time-steps, each corresponding to an hour. The cumulative reward is the summation of all voltage deviations during an episode. The change of reward during the training process is shown in Fig. 2. Before 2000 episodes, the agent actions are randomly chosen to explore the environment and accumulate experience. After that, agents begin to learn, during which the parameters are optimized to maximize the cumulative reward. It can be observed that the cumulative reward keeps increasing and finally converges at about 12000 episodes.

C. Performance Evaluation

To evaluate the performance of the control strategy learned from training data, comparative tests are carried out on test sets, where 30 days of data are used. The average, maximum rise and drop of the voltage deviations, and computing times for various control strategies are shown in Table IV. The original method means that the PV inverters are not controlled. The droop control adopts the QV control strategy. The centralized method assumes perfect communication conditions and the load demand and active power generations of PV are known on time. This strategy provides the theoretical limit of the voltage regulation problem. It can be observed that the MADDPG based control strategy achieves better performance than the droop control method with smaller fluctuations on the voltage profile. Both methods use local information for the decision, however, the MADDPG based method learns the coordinated control strategy by modeling the decision procedure of other agents during training, thus achieves better performance. Compared with the MADDPG-based method, its enhanced version with attention strategy helps the critic of each agent attend to information that is most relevant to the reward during training. Thus, it can further reduce voltage deviation when the number of agents increases.

The average optimization accuracy is defined to evaluate the calculation accuracy of the proposed approach:

$$ACC = \left| \frac{\Delta V_{pro} - \Delta V_{ori}}{\Delta V_{cen} - \Delta V_{ori}} \right| \times 100\% \quad (6)$$

where ACC indicates the optimality of the proposed approach as compared to the theoretical limit; ΔV_{pro}, ΔV_{cen} and ΔV_{ori} are the average voltage deviations of the proposed method, the centralized method and the original value, respectively. The proposed method can reach 95.1% optimality using only local information, demonstrating its effectiveness.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2020.3000652, IEEE Transactions on Power Systems

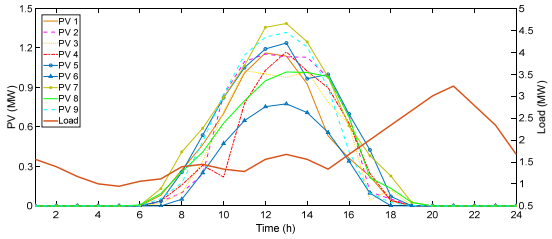

Fig. 3. Profiles of PV generation and load demand.

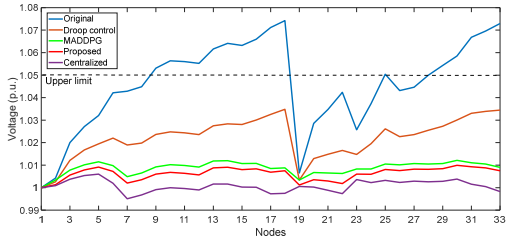

Fig. 4. The voltage of each node before and after optimization when *t=1:00* PM.

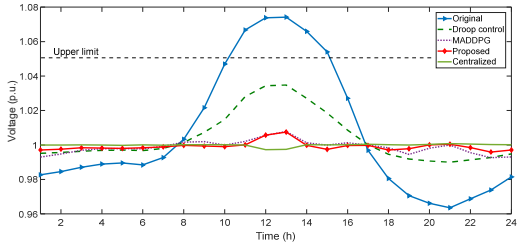

Fig. 5. Voltage change of node 17 before and after optimization.

To further elaborate the performances of each method, a sunny day in the test data is selected. The PV and load profiles are shown in Fig. 3. The objective of this test is to demonstrate the capability of each method in mitigating the over-voltage risk. Comparative results of the voltages at each node before and after reactive power control when *t=1:00 PM* are displayed in Fig. 4. It can be seen that if there are no controls for the PV inverters, there is a concern of the over-voltage problem, see buses 9-18. With the traditional droop control method, that problem can be suppressed. However, compared with our proposed MADDPG method, it has larger voltage fluctuations. The proposed enhancement of MADDPG with attention model further improves the voltage profile. The voltage fluctuations of node 17 across a whole day before and after optimization are shown in Fig. 5. The outcomes are consistent with those observed from Fig. 4, demonstrating an improved performance of the proposed method over other alternatives.

TABLE V Voltage deviation of various methods on test data

Method	Average	Max rise	Max drop	Calculation time (s)
Original	1.89%	7.92%	4.33%	-
Droop	1.26%	3.79%	4.67%	0.0002
MADDPG	0.58%	1.63%	3.85%	0.0012
Proposed	0.53%	1.58%	3.76%	0.0012
Centralized	0.42%	1.52%	3.61%	1.32

D. Test on IEEE 123-bus System

To verify the generality of the proposed approach, simulations are carried out on the IEEE 123-bus system [8]. There are 10 installed PV in total, which are located in nodes 4, 12, 24, 31, 42, 52, 71, 82, 95 and 106, respectively. The performances for various methods are shown in Table V. The results demonstrate that the proposed approach can effectively reduce the voltage deviation of DN. This is consistent with the conclusions on the IEEE 33-bus system.

IV. Conclusions And Future Works

This paper proposes a MADRL based approach with an attention model for distribution system voltage regulation leveraging PV inverters. The proposed method can achieve coordinated control of PV inverters using only local information, thus reduce the cost of communication links. Simulation results on the IEEE 33-bus system demonstrate that the learned coordinated control strategies help better utilize the capability of PV resources and achieve a better control performance. The integration of the attention model further enhances the proposed method to deal with an increased number of control subjects. The proposed method is general and can be easily extended to the PV, wind or other types of DGs integrated systems. Future works will be extending the MADRL based approach for the voltage regulation in unbalance low voltage distribution networks with various types of control devices.

References

[1] Zeraati, Mehdi, Mohamad Esmail Hamedani Golshan, and Josep M. Guerrero. "Distributed control of battery energy storage systems for voltage regulation in distribution networks with high PV penetration." *IEEE Trans. on Smart Grid*, vol. 9, no. 4, pp. 3582-3593, Jul. 2018.

[2] Reinaldo T, Lopes L A C, El-Fouly T H M. "Coordinated active power curtailment of grid connected PV inverters for overvoltage prevention." *IEEE Trans. on Sustain. Energy*, vol. 2, no. 2, pp. 139-147, 2011.

[3] Weckx, Sam, and Johan Driesen. "Optimal local reactive power control by PV inverters." *IEEE Trans. on Sustain. Energy*, vol. 7, no. 4 pp. 1624-1633, 2016.

[4] Stetz, Thomas, Frank Marten, and Martin Braun. "Improved low voltage grid-integration of photovoltaic systems in Germany." *IEEE Trans. on Sustain. Energy*, vol. 4, no. 2, pp. 534-542, 2012.

[5] N. Mahmud, A. Zahedi. "Review of control strategies for voltage regulation of the smart distribution network with high penetration of renewable distributed generation," *Renewable and Sustainable Energy Reviews*, vol. 64, pp. 582-595, Oct. 2016.

[6] H. Ji, C. Wang, *et al*. "A centralized-based method to determine the local voltage control strategies of distributed generator operation in active distribution networks," *Applied Energy*, vol. 228, pp. 2024-2036, Oct. 2018.

[7] K. Baker, A. Bernstein, E. Dall'Anese and C. Zhao, "Network-cognizant voltage droop control for distribution grids," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 2098-2108, Mar. 2018.

[8] H. J. Liu, W. Shi, H. Zhu, "Distributed voltage control in distribution networks: online and robust implementations," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6106-6117, Nov. 2018.

[9] M. Zeraati, M. E. H. Golshan, J. M. Guerrero, "Voltage quality improvement in low voltage distribution networks using reactive power capability of single-phase PV inverters," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5057-5065, Sept. 2019.

[10] S. Y. Wang, J. J. Duan, D. Shi, et al. "A Data-driven multi-agent autonomous voltage control framework using deep reinforcement learning." *IEEE Trans. Power Syst.*, 2020.

[11] R. Lowe, Y. Wu, A. Tamar, *et al*. "Multi-agent actor-critic for mixed cooperative-competitive environments." *Advances in Neural Information Processing Systems*, pp. 6379-6390, 2017.

[12] S. Iqbal, F. Sha, "Actor-attention-critic for multi-agent reinforcement learning." *International Conference on Machine Learning*, pp. 2961-2970, 2019.