Multiagent Twin Delayed Deep Deterministic Policy Gradient Approach for Voltage Control of Distribution System

RIAZ, Hafiz Mehboob, SAJJAD, Malik Intisar Ali and AKMAL, Muhammad <http://orcid.org/0000-0002-3498-4146>

Multiagent Twin Delayed Deep Deterministic Policy Gradient Approach for Voltage Control of Distribution System

Hafiz Mehboob Riaz
Electrical Engineering Department
Univ. of Engg. & Tech. Taxila
Taxila, Pakistan
mehboob.riaz@uettaxila.edu.pk

Malik Intisar Ali Sajjad
Electrical Engineering Department
Univ. of Engg. & Tech. Taxila
Taxila, Pakistan
intisar.ali@uettaxila.edu.pk

Muhammad Akmal
School of Engg. & Built Environment
Sheffield Hallam University
S1 1WB Sheffield, UK
m.Akmal@shu.ac.uk

Abstract— ***Modern power distribution systems face significant challenges, including voltage violations and active power losses, due to the high penetration of renewable energy sources (RESs). Conventional voltage regulation devices are slow and constrained by operational limitations while existing Volt/VAR Control (VVC) techniques for reactive power compensation primarily rely on model-based optimization approaches. In contrast, model-free deep reinforcement learning (DRL) methods, such as Deep Deterministic Policy Gradient (DDPG), can adapt to changing grid conditions. However, DDPG suffers from Q-value overestimation and unstable learning issues, leading to suboptimal control policies. To address these challenges, a Multi-Agent Twin Delayed Deep Deterministic Policy Gradient (MA-TD3) technique is proposed in this paper to optimize the setpoints of modern voltage control devices. By leveraging twin critics with delayed policy updates, MA-TD3 enhances learning stability and mitigates overestimation bias. The distribution network is partitioned into sub-areas, with each sub-area formulated as a Markov game and solved cooperatively using MA-TD3. The proposed approach is validated on a modified IEEE 33-bus system, demonstrating superior performance over existing DRL methods in minimizing voltage violations and active power losses.***

Keywords—Deep Reinforcement Learning (DRL), Multi-Agent systems, Markov Game, Renewable Energy Sources (RESs), Volt-Var Control (VVC)

I. INTRODUCTION

The need to produce clean energy has driven the incorporation of renewable energy sources (RESs) into power distribution systems. However, the widespread integration of these RESs presents substantial challenges e.g., voltage violations and active power losses due to reversed power flow, and intermittency associated with distributed generators (DGs). Conventional devices used to maintain voltage within acceptable ranges in distribution systems include transformers having on-load tap changers (OLTCs) and shunt capacitors (SCs). Although these devices mitigate voltage violations, their mechanical operation results in slow response times and limits their ability to address rapid voltage fluctuations caused by varying DG's power. Additionally, the switching operations of these legacy devices are limited by the distribution system operators (DSOs). Considering the challenges associated with conventional voltage regulating devices, the IEEE 1547 standard [1] allows smart inverters integrated with DGs to exchange reactive power for voltage control, commonly referred to as the volt-VAR control (VVC) problem. Recent approaches to address the VVC problem are generally classified into model-based and data-driven optimization methods [2]. Model-based methods are further divided into classical and heuristic optimization techniques. Classical approaches such as Linear and Quadratic Programming, and Mixed-Integer Programming (MIP) [2], involve continuous and discrete variables. Although classical methods have demonstrated significant performance in achieving the goal of voltage control, they often lack computational efficiency, rendering them unsuitable for real-world grid applications [3]. Heuristic optimization techniques include well-known methods like particle swarm optimization (PSO) and genetic algorithm (GA). While heuristic methods are effective for solving the VVC problem, their performance heavily depends on parameter selection [3]. A common limitation of both classical and heuristic approaches is their reliance on an accurate physical distribution model, which is challenging to achieve in practical scenarios [4].

Data-driven approaches are model-free methods that utilize historical data to address uncertainties in grid models by learning optimal actions. Recent advancements in data-driven methods leverage reinforcement learning (RL), where an agent adapts to dynamic system conditions through continuous interaction with an environment. Several RL-based approaches have been suggested in the literature to regulate the distribution system voltage and reduce active power losses. A Q-learning-based optimal reactive power control strategy has been proposed in [5] to maintain the voltage in the allowed range. A similar Q-learning approach has been proposed in [6] to regulate voltage in distribution systems by optimally determining the tap positions of an OLTC in fluctuating load demands.

However, Q-learning-based methods are limited to only discrete and finite state-action spaces, making them inadequate for real-time voltage control involving continuous actions associated with inverter-based photovoltaics (IBPVs) and SVCs [7]. Deep reinforcement learning (DRL) addresses these limitations by employing deep neural networks as function approximators to extract high-dimensional features while reducing computational burden.

Various DRL approaches have been applied for voltage regulation in distribution systems. A deep Q-network (DQN) based autonomous voltage regulation scheme has been proposed in [8] to adjust generator setpoints under varying load conditions. Another two-level DQN-based VVC technique has been proposed to manage the reactive power of discrete SCs by adjusting their ON/OFF states. While DQN is limited to discrete actions of OLTCs and SCs, DDPG has been widely adopted for continuous control of IBPVs and SVCs. For instance, [9] applies DDPG to adjust IBPV setpoints for instantaneous voltage deviations and capacitor actions for

979-8-3315-6520-6/25/$31.00 ©2025 IEEE

long-term violation mitigation. A multi-agent DDPG scheme has been proposed in [10], [11] to dispatch the conventional and fast voltage controllers on different time scales to regulate the voltage in the distribution system.

The primary limitations of the DDPG algorithm are the unavoidable overestimation error of the Q-value generated by the critic networks and unstable learning, leading to slight deviations from optimal solutions [12]. To mitigate the estimation error in DDPG, the twin delayed deep deterministic policy gradient (TD3) algorithm has been introduced [13], incorporating a clipped double-Q learning with a delayed policy update approach. Recently, TD3 has been applied to address voltage stability in nonlinear DC-DC converters [14] and energy management of electric vehicles [15] . However, its application to voltage control in distribution systems still needs to be explored in detail. This paper proposes a VVC approach for distribution systems using MA-TD3 to optimally determine the setpoints of IBPVs and SVCs. The major contributions of this paper are as follows:

1) To address the Q-value overestimation and unstable learning issues in existing DRL methods, the proposed MA-TD3 uses twin critics with delayed policy update to effectively reduce voltage violation and active power losses in the distribution system.
2) MA-TD3 partitions the distribution system into sub-areas, with agents controlling each area cooperatively to enable coordinated voltage control, reflecting realistic distributed system operations.
3) The approach employs centralized training of different agents with decentralized execution using only local observations, thereby addressing the communication challenges of centralized control architectures.

The rest of the paper is organized as follows: Section II presents the mathematical formulation of voltage regulation devices and the problem formulation of VVC. Section III details the proposed approach. Case studies and result discussion are provided in Section IV.

II. MATHEMATICAL MODELING

This section covers the mathematical modeling of voltage regulating devices e.g., followed by the problem formulation of VVC.

A. Mathematical models of voltage controllers

The primary objective of VVC is to maintain the voltage within predefined limits and minimize the active power losses in the distribution system. This is achieved by determining the optimal reactive power set points of IBPVs and SVCs.

1) Inverter-based PV (IBPV)

The IBPV associated with any bus j can release or absorb reactive power depending on the requirements at every t to mitigate the voltage fluctuations as defined in (1) and (2).

$$Q_j^{IBPV}(t) = \alpha_j^{IBPV}(t).\overline{Q_j^{IBPV}}(t) \tag{1}$$

$$Q_j^{IBPV}(t) \le \sqrt{\left(S_j^{IBPV}(t)\right)^2 - \left(P_j^{IBPV}(t)\right)^2} \tag{2}$$

where, $\overline{Q_j^{IBPV}}(t)$ is the maximum reactive power delivered by the IBPV while $S_j^{IBPV}(t)$ and $P_j^{IBPV}(t)$ represent the apparent and active power from IBPV respectively. The parameter $\alpha_j^{IBPV}(t)$ is defined in the range $\alpha_j^{IBPV}(t) \in [-1,1]$.

2) Static VAR compensator (SVC)

SVC belongs to a family of flexible AC transmission (FACTS) devices and is capable of continuously exchanging reactive power to realize VVC. The reactive power of SVC Q_j^{SVC} is defined in (3) as:

$$\underline{Q_j^{SVC}}(t) \le Q_j^{SVC}(t) \le \overline{Q_j^{SVC}}(t) \tag{3}$$

where $\underline{Q_j^{SVC}}(t)$ and $\overline{Q_j^{SVC}}(t)$ represent the lower and upper limit of reactive power supplied by the SVC.

B. VVC problem Formulation

A radial distribution system is generally represented with a set $\mathbb{N} = \{1,2..N_b\}$ where N_b is the total number of buses. The power flow equations of the branches are given by the Distflow equation as follows;

$$\min_{\substack{Q_{IBPV}(t)\\ Q_{SVC}(t)}} \sum_{t=1}^{N_t} \sum_{i=1}^{N_b} \left[(v_i(t)-V_0)^2 + r_i \frac{P_i^2(t)+Q_i^2(t)}{v_i^2(t)} \right] \tag{4}$$

Subject to:

$$\sum_{i\in m(j)} \left[P_{ij}(t) - i_{ij}^2(t) r_{ij}\right] - P_j(t) = \sum_{k\in n(j)} P_{jk}(t) \tag{5}$$

$$\sum_{i\in m(j)} \left[Q_{ij}(t) - i_{ij}^2(t) x_{ij}\right] - Q_j(t) = \sum_{k\in n(j)} Q_{jk}(t) \tag{6}$$

$$v_j^2(t) = v_i^2(t) + \left(r_{ij}^2 + x_{ij}^2\right) i_{ij}^2(t) - 2\left(r_{ij}P_{ij}(t) + x_{ij}Q_{ij}(t)\right) \tag{7}$$

$$P_{ij}^2(t) + Q_{ij}^2(t) = \left(i_{ij}(t) v_i(t)\right)^2 \tag{8}$$

$$P_j(t) = P_j^l(t) - P_j^{IBPV}(t) \tag{9}$$

$$Q_j(t) = Q_j^l(t) - Q_j^{IBPV}(t) - Q_j^{SVC}(t) \tag{10}$$

$$\underline{v_i}(t) \le v_i(t) \le \overline{v_i}(t) \tag{11}$$

where, the objective function in (4) defines the minimization problem of the sum of voltage violation and active power losses in distribution system; (5)-(6) ensure active & reactive power balance at each node with $m(j)$ and $n(j)$ as the parent and child bus set of node j; P_{ij}, Q_{ij} are the active and reactive power flows from bus i to j; r_{ij}, x_{ij} are the resistance and reactance of the line segment (i, j); (7) models the voltage drop along (i, j); (8) relates squared branch current to power flows and sending-end voltage; (9)–(10) define the net active & reactive power injections at bus j; P_j^l, Q_j^l are load's active & reactive power; (11) bounds each bus voltage magnitude within the upper $\overline{v_i}(t)$ and lower limit $\underline{v_i}(t)$. The minimization problem in (4) is inherently non-convex and NP-hard due to nonlinear power and voltage relationship in (8). Given the dynamic and reconfigurable nature of distribution systems, model-based VVC solutions often yield suboptimal results. Therefore, a model-free DRL-based solution for the VVC problem is proposed in the next section.

III. PROPOSED MA-TD3 BASED VVC

This section covers the mathematical formulation of VVC problem in the form of Markov games, followed by the proposed multi-agent DRL technique to solve the VVC problem.

A. Formulation of Markov game

To address the VVC problem using a multi-agent DRL approach, the distribution system is partitioned into sub-areas based on voltage and reactive power sensitivity, with each sub-area managed by a voltage controller, acting as an agent. By operating cooperatively, all agents work together to achieve VVC objectives. The agents' optimal set points are formulated as a Markov game, and are defined as a tuple, $([\boldsymbol{O}_b]_n, \boldsymbol{S}, [\boldsymbol{A}_b]_n, \boldsymbol{R}, \boldsymbol{P})$ in which the variable $s \in \boldsymbol{S}$ comprise global state of system , $O_t^b \in \boldsymbol{O}_b$ is the local observation of an agent at time t belonging to area b, $a_t^b \in \boldsymbol{A}_b$ is the agent b's action, while the reward of all the agents is $\boldsymbol{R} \rightarrow \boldsymbol{S} \times \boldsymbol{A}_1 \times \boldsymbol{A}_2 \ldots \boldsymbol{A}_n$ and $\boldsymbol{P}$ denotes the transition probability: $\boldsymbol{S} \times \boldsymbol{A}_1 \times \boldsymbol{A}_2 \ldots \boldsymbol{A}_n \times \boldsymbol{S} \rightarrow [0,1]$. Whenever an agent observes a state O_t^b at any time t, selects an action a_t^b based on its policy $\boldsymbol{\pi}: \boldsymbol{S} \rightarrow \boldsymbol{A}$, gets a reward $\boldsymbol{R}$.

1) State

The local observation of an agent at time t is $\boldsymbol{o}_b(t)=[\boldsymbol{V}^T, \boldsymbol{P}_g^T, \boldsymbol{P}_{Load}^T, \boldsymbol{Q}_g^T]^T$ where $\boldsymbol{P}_g$, $\boldsymbol{Q}_g$ are the vectors of active and reactive power injection, $\boldsymbol{V}$ is the vector of voltage magnitude at all the nodes and $\boldsymbol{P}_{Load}^T$ indicates the load's active power in area b.

2) Action

The action set of IBPV and SVCs agents are defined as: $\boldsymbol{A}_{IBPV}(t)=[\boldsymbol{a}_{IBPV1}(t), \boldsymbol{a}_{IBPV2}(t), \ldots \boldsymbol{a}_{IBPVN}(t)]^T$, $\boldsymbol{A}_{SVC}(t)=[\boldsymbol{a}_{SVC1}(t), \boldsymbol{a}_{SVC2}(t), \ldots \boldsymbol{a}_{SVCN}(t)]^T$ with $\boldsymbol{A}_{SVC}(t)=\boldsymbol{A}_{IBPV}(t)=Q_g \in [-1,1]$ representing that action will inject or absorb the reactive power to attain the objectives of VVC.

3) Reward

The VVC problem is generally expressed as a multi-objective function, for the minimization of voltage violations and active power losses. Thus, the reward function of an agent b is a weighted sum of these two terms, defined in (12) as:

$$r_b(t) = r_p(t) + c\, r_v(t) \tag{12}$$

$$r_p(t) = -\sum_{i=0}^{N_B} P_{loss}(t) \tag{13}$$

$$r_v(t) = -\sum_{i=0}^{N_B} \left[max\,(v_i(t) - \overline{v_i}(t),\, 0) + max\,(\underline{v_i}(t) - v_i(t),\, 0) \right] \tag{14}$$

The voltage violation term in the reward function is penalized using a reward scaling strategy adopted from [16], highlighting that learning voltage violation rewards is more challenging.

B. Proposed MA-DRL scheme for VVC

To solve the formulated VVC problem, MA-TD3 framework involves centralized training followed by decentralized execution. Each agent is associated with an actor and two critic networks. The actor selects an action based on its local observation, while the critic evaluates the actor's actions by using the global state to compute the state-action value, also known as Q-value. The expected cumulative future reward attained by all the agents is termed as Q-function defined in (15) as:

$$Q^{\pi}(s, a_t^1 \ldots a_t^n) = \underset{a_b \sim \pi_b}{E} \sum_{t=0}^{\infty} (\gamma^t r_t | s, a_t^1 \ldots a_t^n) \tag{15}$$

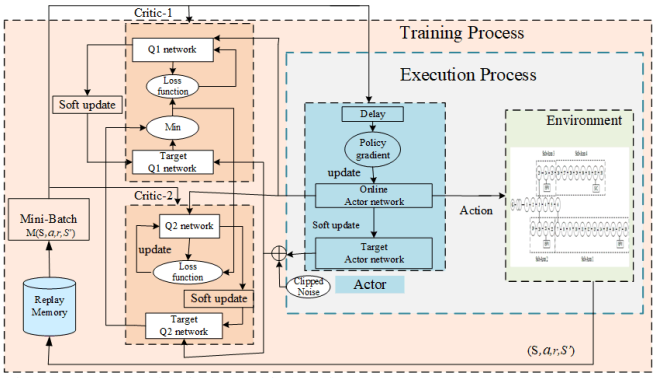

Fig. 1. Framework of proposed TD3 for VVC

DRL aims to maximize its cumulative reward by finding an optimal policy π^*, defined in (16) as:

$$\pi^* = \underset{\pi_b}{argmax}\, E_{\pi}\, Q^{\pi}(s, a_b) \tag{16}$$

In practical implementation, actor and critic networks are realized by deep neural networks (DNNs). The critic network Q_{ψ_b} learns the state-action value function using the Bellman equation defined in (17) as:

$$y_t = \underset{a_b \sim \pi_b}{E} \left[r_{b,t} + \gamma\, Q_{\psi_b}(s_{t+1}, a_{t+1}^1 \ldots a_{t+1}^n | s_t, a_t) \right] \tag{17}$$

where, y_t is the estimated value of $Q_{\psi_b}(s_t, a_t)$ from the last iteration. The critic network learns to minimize the loss function defined in (18):

$$L_b = \underset{\psi_b}{arg\,min} \left(Q_{\psi_b}(s_t, a_t^1 \ldots a_t^n) - y_t \right)^2 \tag{18}$$

The actor-network adjusts the parameters θ_b in the direction of maximizing the objective function defined in (19) as:

$$\nabla_{\theta_b} J(\theta_b) = \underset{s_t,\, a_t \sim M}{E} \left[\nabla_{\theta_b} \pi_{\theta_b}(a_t^b | s_t^b)\, \nabla_{a_t^b} Q_b^{\pi}(s_t, a_t^1 \ldots a_t^n) \right] \tag{19}$$

The critic function is optimized by minimizing the difference between the estimated Q-value and its target value. However, training instability arises as the same critic being updated is also used to compute the target value. To address this, copies of actor and critic networks, termed target actor and target critic networks with parameters π'_{θ} and Q'_{ψ} are created. These networks track the online critic using a soft update mechanism defined in (20)-(22) as:

$$\pi'_{\theta} \leftarrow \omega \pi_{\theta} + (1-\omega)\pi'_{\theta} \tag{20}$$

$$Q'_{\psi_1} \leftarrow \omega Q_{\psi_1} + (1-\omega) Q'_{\psi_1} \tag{21}$$

$$Q'_{\psi_2} \leftarrow \omega Q_{\psi_2} + (1-\omega) Q'_{\psi_2} \tag{22}$$

where, $\omega \ll 1$ is used to update the target networks. Equation (17) is re-defined considering target networks as:

$$y_t = \underset{a_b \sim \pi_b}{E}\left[r_{b,t} + \gamma\, Q'_{\psi_b}\left(s_{t+1}, a^1_{t+1} \ldots a^n_{t+1} | s_t, a_{,t}\right)\right] \quad (23)$$

To mitigate overestimation bias arising from function approximation errors, TD3 employs two critic networks and uses the minimum of their outputs to compute the target value y_t, as defined in (24).

$$y_t = \underset{a_b \sim \pi_b}{E}\left[r_{b,t} + \gamma \min_{n=1,2} Q'_{\psi_{b,n}}\left(s_{t+1}, \widehat{a^1_{t+1}} \ldots \widehat{a^n_{t+1}} | s_t, a_t\right)\right] \quad (24)$$

Here, the minimum of two target critics $Q'_{\psi_{b,n}}$ evaluates next step joint actions $\widehat{a^1_{t+1}}$ sampled from the agents' target policies, conditioned on current state s_t and action a_t.

To alleviate the overfitting problem and smooth target policy, Gaussian noise $N(0, \sigma)$ is incorporated to the target action $\widehat{a^1_{t+1}}$, and clipped within the range $[\underline{\xi}, \bar{\xi}]$ as defined in (25)

$$\widehat{a^1_{t+1}} = \widehat{a^1_{t+1}} + clip\left(N(0, \sigma), \underline{\xi}, \bar{\xi}\right) \quad (25)$$

During execution, actions are taken without noise, while clipping enforces reactive power limits. A replay buffer stores the agent's interaction with the environment, from which M mini-batch samples are drawn to train actor and critic networks. The use of a replay buffer requires redefining the optimization of actor and critic parameters. The critic networks learn by minimizing mean square error loss (MSE) defined in (26) as:

$$L_{Q_\psi} = \frac{1}{|M|} \sum_{(s, a_1 \ldots a_n, r_{pl}) \in M} \left(Q_\psi(s_t, a^1_t \ldots a^n_t) - y_t\right)^2 \quad (26)$$

The parameter update mechanism for ψ uses the gradient descent method as:

$$\psi \leftarrow \psi - \lambda_Q \nabla_\psi L_{Q_\psi} \quad (27)$$

Here, λ_Q is the critic network's learning rate. The actor networks are updated in (28) using the gradient ascent method to maximize their respective objective functions.

$$\nabla_{\Theta_b} J(\Theta_b) = \frac{1}{|M|} \sum_{(s, a_1 \ldots a_n, r_{pl}) \in M} \nabla_{\Theta_b} \pi_{\Theta_b}(a^b_t | s^b_t)\, \nabla_{a^b_t} Q^\pi_b(s_t, a^1_t \ldots a^n_t)^2 \quad (28)$$

Where, the first term is the gradient of the policy output with respect to its parameters, and the second term is the gradient of the centralized critic with respect to agent b's action, capturing the joint impact of all agents' actions in state s_t. The updated mechanism of actor networks' parameters $\theta_1, \ldots \theta_n$ is given in (29) as:

$$\Theta \leftarrow \Theta + \lambda_\pi \nabla_{\Theta_b} J(\Theta_b) \quad (29)$$

The details of the proposed approach are summarized in the algorithm presented in Table 1 while the framework is illustrated in Fig. 1.

TABLE I THE PROPOSED MA-TD3 ALGORITHM FOR VVC

for each agent IBPVs, SVCs **do**
  Randomly initialize actor-network's parameter θ_π and critic network ψ_{Q1}, ψ_{Q2} and empty replay memory D;
end for
  Initialize parameters of target actor and critic networks $\psi'_{Q1} \leftarrow \psi_{Q1}$, $\psi'_{Q2} \leftarrow \psi_{Q2}$, $\theta'_\pi \leftarrow \theta_\pi$ for each agent
  for $t = 1$ *to* T **do**
    Get the initial state s_t for each agent b
    Select the action using (28), execute the action $a^1_t \ldots a^n_t$ in the environment, collect reward $r_b(t)$, and new state s_{t+1}
    Store $(s_t, a^1_t \ldots a^n_t, r_b(t), s_{t+1})$ in D
    for agent b = 1…N, **do**
      sample a random mini-batch transition from D
      find target y_t based on (24)
      update critic networks using (26-27)
      if t mod policy update frequency = 0 then
        update actor network using (28-29)
        soft update target networks using (20-22)
    end for
  end for

IV. CASE STUDIES AND RESULT DISCUSSION

This section validates the effectiveness of the proposed MA-TD3 approach to solve VVC problem for IEEE 33-bus system. The load and generation data are sourced from eastern China [17]. The proposed DRL algorithm is implemented using the PyTorch framework, while the modeling of test system and balanced power flow calculations are conducted with Pandapower [18]. The test system data is obtained from Matpower [19]. To train the DRL algorithm, data from 300 days is utilized, where the generation, as well as load levels have been scaled according to the daily fluctuation ratio. To alleviate the impact of randomness, the proposed method is evaluated using three distinct random seeds, and average results are presented.

These seeds influence neural network weight initialization and action exploration noise. This multi-seed evaluation aims to ensure the learned voltage control policy's robustness. The parameters of the proposed MA-TD3 algorithm are detailed in Table II.

The effectiveness of the proposed algorithm is evaluated against two state-of-the-art DRL algorithms: multi-agent DDPG and multi-agent SAC.

Fig 2 illustrates the network topology of the modified IEEE 33-bus system, where the test system has been divided into four sub-areas using the shortest route starting from the terminal to the main branch comprising the nodes 1-6 [20], and each sub-area is controlled by an agent. In sub-areas one, two, and three, total three IBPVs are installed at buses 17, 21, and 24 respectively, each with a capacity of 1.5 MW active and 2 MVAR reactive power. Additionally, one SVC with a capacity of 2 MVAR reactive power is connected at bus 32.

Fig. 3 presents the testing results during the training phase on the IEEE 33-bus system, showing daily accumulated reward, active power loss, and voltage violation rate over episodes.

TABLE II PARAMETER SETTING FOR THE PROPOSED ALGORITHM

Parameter	Value
Activation Function	ReLU
Optimizer	Adam
Hidden layers	2
Policy update frequency	2
Replay memory size	30000
Mini-batch size	128
Actor-network's learning rate	0.0001
Critic-network's learning rate	0.0003
Coefficient of voltage violation	50
Target policy smoothing noise	0.2
Discount factor	0.90
Soft update parameter	0.001
Exploration noise	$N(0, 0.05)$
Neurons for actor & critic hidden layer	512

The training phase was tested using three independent random seeds to ensure consistency and robustness.

The proposed MA-TD3 algorithm demonstrates superior performance across all metrics, outperforming the baseline algorithms MA-DDPG and MA-SAC.

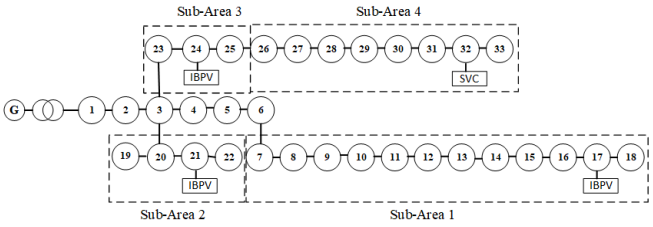

Fig. 2. The modified IEEE-33 bus topology

In the cumulative reward plot, MA-TD3 achieves the highest cumulative reward (2.32), with a remarkably fast convergence, indicating the agent's ability to rapidly learn an effective control policy. This high reward value reflects the algorithm's efficient balance between minimizing voltage violations and reducing power loss, since the reward function combines both objectives. In contrast, MA-DDPG (-3.37) and MA-SAC (-2.96) settle at lower reward levels and exhibit slower convergence, highlighting their relative inefficiency in multi-objective optimization.

The active power loss plot further reinforces the effectiveness of MA-TD3. It converges quickly to the lowest steady-state value (2.28 p.u.), outperforming MA-DDPG (3.21 p.u.) and MA-SAC (2.89 p.u.). MATD3's smoother trajectory with minimal oscillations reflects greater training stability and energy efficiency, which are essential for reliable grid operation. This improvement is largely due to TD3's architectural features, including twin critics to reduce Q-value overestimation and delayed policy updates contribute to more stable and reliable learning dynamics.

In the voltage violation rate plot, MA-TD3 again demonstrates superior performance, achieving a sharp decline in violations by episode 10, which is faster than MA-DDPG (11th episode) and MA-SAC (12th episode). Moreover, the steady-state voltage violation rate of MA-TD3 is the lowest (1.0337e−7), compared to MA-DDPG (2.483e−7) and MA-SAC (1.509e−7). The zoomed-in insets further reveal that TD3 maintains an extremely low violation rate throughout training, with minimal transient spikes, while the other methods show frequent deviations, suggesting less robust voltage regulation behavior.

These observations are further validated in Table III. The proposed MA-TD3 algorithm achieves the highest final reward (−2.32), lowest power loss (2.28 p.u.), and the least voltage violation in the final 50 episodes, confirming its robustness and superiority over the baseline algorithms. Overall, MA-TD3 not only ensures faster convergence and stable learning but also delivers the most optimal performance in Volt/VAR control for active distribution networks.

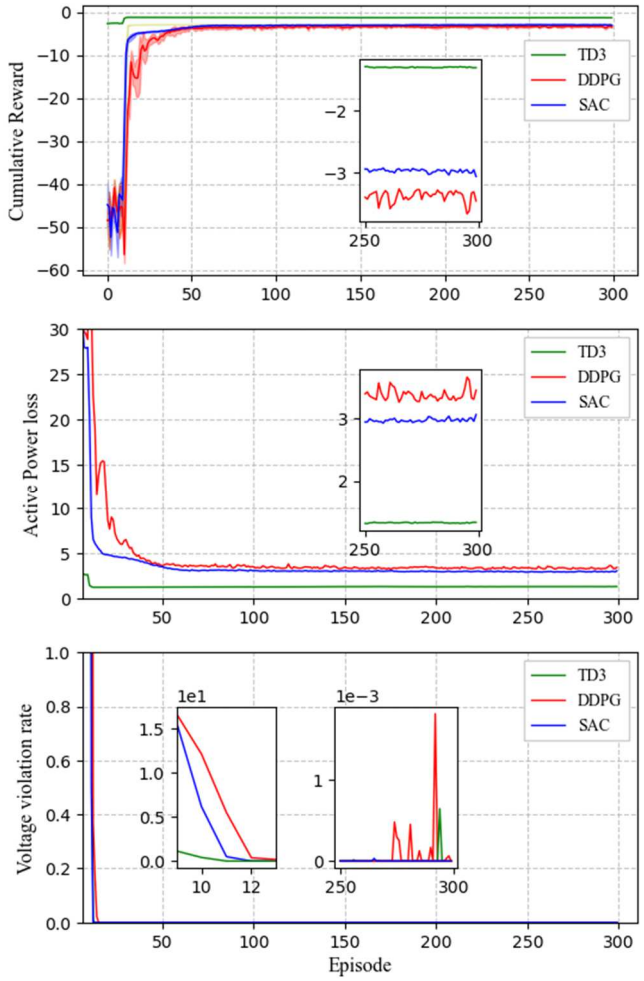

Fig. 3. Testing outcomes during training for IEEE 33-bus system

TABLE III PERFORMANCE EVALUATION AGAINST DRL METHODS

Algorithm	Performance indices		
	P_{loss}/MW	VVR/p.u	Reward
MA-DDPG	3.210	2.483e-7	-3.371
MA-SAC	2.897	1.509e-7	-2.967
Proposed	**2.280**	**1.037e-7**	**-2.320**

V. CONCLUSION

This paper addresses the challenges of voltage violations and active power losses in modern power distribution systems. Conventional VVC methods, constrained by operational limitations, often rely on model-based optimization, while DRL-based approaches like DDPG struggle with unstable learning issues and Q-value overestimation. To overcome

these limitations, we proposed a MA-TD3 approach for optimal voltage control. By incorporating twin critics and delayed policy updates, the proposed method enhances learning stability, enabling decentralized yet cooperative coordination of voltage regulation devices.

Simulations on the IEEE 33-bus system show that MA-TD3 achieves faster convergence, higher rewards, and superior control performance, significantly reducing power loss (2.280 p.u.) and maintaining a low voltage violation rate with minimal transients. These results demonstrate MA-TD3's robustness and effectiveness for scalable and real-world voltage control applications.

Future work will extend the proposed method to larger distribution systems and incorporate discrete action voltage control devices such as OLTC and SCs to enhance voltage support and overall system flexibility.

REFERENCES

[1] *2019 IEEE 46th Photovoltaic Specialists Conference (PVSC)*. IEEE, 2019.

[2] S. M. Abdelkader *et al.*, "Advancements in data-driven voltage control in active distribution networks: A Comprehensive review," Sep. 01, 2024, *Elsevier B.V.* doi: 10.1016/j.rineng.2024.102741.

[3] H. Mataifa, S. Krishnamurthy, and C. Kriger, "Volt/VAR Optimization: A Survey of Classical and Heuristic Optimization Methods," *IEEE Access*, vol. 10, pp. 13379–13399, 2022, doi: 10.1109/ACCESS.2022.3146366.

[4] T. Xu, W. Wu, Y. Hong, J. Yu, and F. Zhang, "Data-driven Inverter-based Volt/VAr Control for Partially Observable Distribution Networks," *CSEE Journal of Power and Energy Systems*, vol. 9, no. 2, pp. 548–560, Mar. 2023, doi: 10.17775/CSEEJPES.2020.05920.

[5] Y. Xu, W. Zhang, W. Liu, and F. Ferrese, "Multiagent-based reinforcement learning for optimal reactive power dispatch," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 42, no. 6, pp. 1742–1751, 2012, doi: 10.1109/TSMCC.2012.2218596.

[6] H. Xu, A. D. Dominguez-Garcia, and P. W. Sauer, "Optimal Tap Setting of Voltage Regulation Transformers Using Batch Reinforcement Learning," *IEEE Transactions on Power Systems*, vol. 35, no. 3, pp. 1990–2001, May 2020, doi: 10.1109/TPWRS.2019.2948132.

[7] Y. Wang, V. Vittal, A. Pal, and M. Hedman, "Deep Reinforcement Learning Based Voltage Controls for Power Systems under Disturbances," 2024.

[8] *2019 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2019.

[9] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-Timescale Voltage Control in Distribution Grids Using Deep Reinforcement Learning," *IEEE Trans Smart Grid*, vol. 11, no. 3, pp. 2313–2323, May 2020, doi: 10.1109/TSG.2019.2951769.

[10] X. Sun and J. Qiu, "Two-Stage Volt/Var Control in Active Distribution Networks with Multi-Agent Deep Reinforcement Learning Method," *IEEE Trans Smart Grid*, vol. 12, no. 4, pp. 2903–2912, Jul. 2021, doi: 10.1109/TSG.2021.3052998.

[11] Z. Wu *et al.*, "Multi-timescale voltage control for distribution system based on multi-agent deep reinforcement learning," *International Journal of Electrical Power and Energy Systems*, vol. 147, May 2023, doi: 10.1016/j.ijepes.2022.108830.

[12] S. Fujimoto, D. Meger, and D. Precup, "Off-Policy Deep Reinforcement Learning without Exploration." [Online]. Available: https://github.com/sfujim/BCQ

[13] S. Fujimoto, H. Van Hoof, and D. Meger, "Addressing Function Approximation Error in Actor-Critic Methods," 2018. [Online]. Available: https://github.com/

[14] A. Rajamallaiah, S. P. K. Karri, and Y. R. Shankar, "Deep Reinforcement Learning Based Control Strategy for Voltage Regulation of DC-DC Buck Converter Feeding CPLs in DC Microgrid," *IEEE Access*, vol. 12, pp. 17419–17430, 2024, doi: 10.1109/ACCESS.2024.3358412.

[15] B. Chen *et al.*, "A hierarchical cooperative eco-driving and energy management strategy of hybrid electric vehicle based on improved TD3 with multi-experience," *Energy Convers Manag*, vol. 326, p. 119508, 2025, doi: https://doi.org/10.1016/j.enconman.2025.119508.

[16] Q. Liu *et al.*, "Two-Critic Deep Reinforcement Learning for Inverter-Based Volt-Var Control in Active Distribution Networks," *IEEE Trans Sustain Energy*, vol. 15, no. 3, pp. 1768–1781, Jul. 2024, doi: 10.1109/TSTE.2024.3376369.

[17] H. Liu and W. Wu, "Two-Stage Deep Reinforcement Learning for Inverter-Based Volt-VAR Control in Active Distribution Networks," *IEEE Trans Smart Grid*, vol. 12, no. 3, pp. 2037–2047, May 2021, doi: 10.1109/TSG.2020.3041620.

[18] L. Thurner *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, Nov. 2018, doi: 10.1109/TPWRS.2018.2829021.

[19] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12–19, Feb. 2011, doi: 10.1109/TPWRS.2010.2051168.

[20] J. Wang, W. Xu, Y. Gu, W. Song, and T. C. Green, "Multi-Agent Reinforcement Learning for Active Voltage Control on Power Distribution Networks," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2110.14300