International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 744-757, May-2026

Entropy-Regularized Deep Reinforcement Learning for Stochastic Voltage Regulation under High Renewable Penetration

Tarun Kumar Modi*, Naresh Sapate*, Shailendra Turkar*
*Department of Electrical Engineering
Sardar Patel University, Balaghat, India
Email: tarunk24x7@gmail.com

***Abstract*—The rapid growth of renewable energy sources, particularly solar photovoltaic and wind generation, has fundamentally changed the operating characteristics of modern power distribution networks. The inherent variability and limited predictability of these resources create substantial challenges for voltage regulation, as traditional control schemes struggle to respond effectively to fast and unpredictable fluctuations. This paper presents an entropy-regularized deep reinforcement learning framework designed specifically for stochastic voltage regulation in distribution grids with high renewable penetration. Unlike conventional reinforcement learning methods that converge to deterministic policies, the proposed approach maintains policy stochasticity through entropy regularization, which encourages exploration and improves robustness against the uncertainty introduced by renewable generation. We develop a Soft Actor-Critic based control agent that coordinates reactive power from smart inverters, on-load tap changers, and static var compensators to maintain voltage within acceptable bounds while accounting for the probabilistic nature of renewable output. The framework is validated through extensive simulations on a modified IEEE 33-bus test system with 65% renewable penetration. Results demonstrate that the entropy-regularized approach reduces voltage violations by 87% compared to rule-based control and achieves 23% better performance than standard deep reinforcement learning methods under highly variable generation conditions. The proposed method also exhibits superior generalization when tested on unseen scenarios with different renewable generation patterns.**

***Index Terms*—Deep reinforcement learning, entropy regularization, voltage regulation, renewable energy, distribution networks, soft actor-critic, stochastic control**

I. Introduction

A. The Changing Landscape of Distribution Networks

Over the past decade, power distribution networks have undergone a remarkable transformation. What were once passive systems designed simply to deliver electricity from transmission substations to end consumers have evolved into active networks hosting substantial amounts of distributed generation. Solar panels on rooftops, community wind turbines, and battery storage systems now inject power directly into medium and low voltage grids, creating operational patterns that network designers never anticipated.

This transformation brings undeniable benefits for sustainability and energy independence, but it also introduces complications that grid operators must learn to manage. Among these, voltage regulation stands out as particularly challenging. When clouds pass over a neighborhood with high solar adoption, generation can drop by 70% or more within seconds. When the sun emerges again, voltage can spike rapidly as power flows reverse direction. These dynamics repeat countless times throughout the day, stressing equipment and threatening power quality for all connected customers.

Traditional voltage regulation equipment was never designed for such conditions. Mechanical tap changers operate on timescales of seconds to minutes, far too slow to track cloud-induced ramps. Capacitor banks switch in discrete steps, unable to provide the continuous adjustment that variable generation demands. Even modern smart inverters,

while capable of fast response, typically operate based on local measurements without awareness of conditions elsewhere in the network.

B. Why Stochastic Control Matters

The fundamental difficulty with renewable generation lies not in its variability per se, but in its unpredictability. A gas turbine varies its output too, but it does so according to dispatch commands that the operator controls. Solar and wind generation vary according to weather, which no one controls and no one predicts perfectly. This distinction has profound implications for control system design.

Consider a voltage controller that learns to expect certain generation patterns based on historical data. Such a controller may perform beautifully on typical days but fail dramatically when an unusual weather pattern produces generation profiles it has never encountered. This brittleness represents a serious practical concern, because the worst voltage problems tend to occur precisely during unusual conditions—unexpected heat waves that drive up air conditioning load while reducing solar panel efficiency, or sudden storms that simultaneously cut solar output and increase wind generation.

What we need, then, is a control approach that explicitly acknowledges uncertainty and prepares for it. Rather than learning a single ”best” response to each situation, the controller should maintain flexibility, keeping multiple viable options available and selecting among them based on real-time conditions. This is exactly what entropy-regularized reinforcement learning provides.

C. Contributions of This Work

This paper develops an entropy-regularized deep reinforcement learning framework for voltage regulation in distribution networks with high renewable penetration. The work makes several distinct contributions to the field.

First, we formulate the voltage regulation problem in a way that explicitly captures the stochastic nature of renewable generation. Rather than treating variability as noise to be filtered out, we model it as a fundamental characteristic of the environment that the controller must learn to handle.

Second, we apply entropy regularization to the reinforcement learning objective, which encourages the learned policy to maintain beneficial stochasticity. This improves both exploration during training and robustness during deployment, as the controller naturally hedges against uncertainty rather than committing fully to any single action.

Third, we develop a practical implementation based on the Soft Actor-Critic algorithm, adapted for the specific requirements of distribution network control. This includes appropriate state representations, action spaces, and reward functions that reflect actual operational objectives.

Fourth, we conduct comprehensive validation studies that examine not only average performance but also behavior under extreme conditions and generalization to scenarios not seen during training. These aspects are critical for practical deployment but often overlooked in academic studies.

II. Background and Related Work

A. Classical Voltage Regulation

Voltage regulation in distribution networks has a long history, with techniques evolving alongside the networks themselves. The most established approach uses on-load tap changers at distribution transformers, which adjust the turns ratio to compensate for voltage drops along feeders. These devices typically operate based on voltage measured at the substation or at a remote point, with deadbands and time delays to prevent excessive operations.

Shunt capacitors provide another traditional tool, compensating for reactive power consumption by loads and thereby reducing voltage drops. Originally installed as fixed compensation, modern capacitor banks often include switching capability to adjust compensation based on loading conditions. However, their discrete nature limits the precision of voltage control they can achieve.

Step voltage regulators, essentially autotransformers with tap-changing capability, offer additional flexibility by allowing voltage adjustment at points along the feeder rather than only at the substation. Distribution network operators have traditionally used these devices in combination, establishing coordinated control schemes that balance voltage quality against equipment wear.

These classical approaches work well enough when load and generation are predictable. The fundamental assumption underlying their design is

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 744-757, May-2026

that conditions change slowly enough for human operators or simple automation to respond appropriately. High renewable penetration violates this assumption, creating needs that classical equipment cannot fully address.

B. Inverter-Based Voltage Support

The proliferation of power electronic inverters for renewable generation and energy storage has created new possibilities for voltage regulation. Modern inverters can adjust their reactive power output continuously and nearly instantaneously, providing a fast-acting complement to slower mechanical devices.

The IEEE 1547-2018 standard now requires distributed energy resource inverters to provide voltage support functions, including Volt-VAR and Volt-Watt modes. In Volt-VAR mode, inverters automatically absorb or inject reactive power based on local voltage measurements, following a configurable characteristic curve. Volt-Watt mode curtails active power output when voltage rises too high, providing a last resort for overvoltage prevention.

These autonomous functions represent a significant advance, but they have limitations. Because each inverter responds only to its local voltage, there is no guarantee that the collective behavior will be optimal or even stable. Studies have documented cases of oscillations and adverse interactions when multiple inverters with aggressive settings operate in proximity. Furthermore, purely local control cannot anticipate problems—it only reacts after voltage has already deviated.

Coordinated inverter control schemes attempt to address these shortcomings through communication and centralized optimization. However, such schemes require reliable communications infrastructure and substantial computational resources, and they may struggle with the latency inherent in solving optimization problems for real-time control.

C. Reinforcement Learning for Power Systems

Reinforcement learning has attracted considerable attention in recent years as a potential solution for complex power system control problems. The appeal is understandable: reinforcement learning agents can learn effective control policies directly from interaction with the system, without requiring explicit models of system dynamics.

Early work applied tabular reinforcement learning methods to relatively simple power system problems, such as optimal capacitor switching and demand response. As deep learning advanced, researchers began combining neural networks with reinforcement learning to handle the high-dimensional state spaces characteristic of realistic power system models.

For voltage control specifically, several research groups have explored deep reinforcement learning approaches. Some have focused on training individual inverter controllers, while others have developed centralized agents that coordinate multiple devices. Both deterministic policy methods, such as Deep Deterministic Policy Gradient, and stochastic policy methods, such as Proximal Policy Optimization, have been applied.

However, much of this prior work has treated the reinforcement learning problem in a relatively standard way, without specifically accounting for the stochastic nature of renewable generation. The resulting policies may perform well on average but prove fragile when conditions deviate from the training distribution. This limitation motivates our focus on entropy regularization, which specifically promotes robust behavior in uncertain environments.

D. Entropy Regularization in Reinforcement Learning

The idea of adding entropy to the reinforcement learning objective dates back several decades, but it has gained renewed prominence through recent algorithmic developments. The basic concept is simple: rather than maximizing expected reward alone, the agent maximizes expected reward plus a weighted entropy bonus that encourages stochastic action selection.

This formulation offers several advantages. During training, the entropy bonus prevents premature convergence to suboptimal deterministic policies by maintaining exploration. During deployment, the stochastic policy provides natural robustness against model mismatch and distributional shift, because the agent does not commit fully to any single action.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 744-757, May-2026

The Soft Actor-Critic algorithm, developed at UC Berkeley, provides an efficient and stable implementation of entropy-regularized reinforcement learning for continuous action spaces. It has demonstrated strong performance across a range of robotic control tasks and has begun to see application in other domains. To our knowledge, however, it has not been systematically applied to voltage regulation in distribution networks with high renewable penetration.

III. Problem Formulation

A. Distribution Network Model

We consider a radial distribution network with $N+1$ buses, indexed from 0 to N, where bus 0 represents the substation. The network topology is described by a tree graph in which each bus except the substation has exactly one upstream neighbor. We denote the unique path from the substation to bus i as $\mathcal{P}_i$.

At each bus i, we define the complex power injection as $S_i = P_i + jQ_i$, where positive values indicate injection into the network and negative values indicate consumption. For buses with renewable generation, the active power injection includes a stochastic component:

$$P_i^{gen}(t) = \bar{P}_i^{gen}(t) + \xi_i(t) \tag{1}$$

where $\bar{P}_i^{gen}(t)$ represents the expected generation based on forecasts and $\xi_i(t)$ captures the forecast error. We model $\xi_i(t)$ as a zero-mean stochastic process whose statistical properties depend on the type and capacity of generation at bus i.

The relationship between power flows and voltages is governed by the AC power flow equations. For radial networks, these can be expressed efficiently using the DistFlow formulation:

$$P_{ij} = P_j + r_{ij}\frac{P_{ij}^2 + Q_{ij}^2}{V_i^2} + \sum_{k \in \mathcal{C}_j} P_{jk} \tag{2}$$

$$Q_{ij} = Q_j + x_{ij}\frac{P_{ij}^2 + Q_{ij}^2}{V_i^2} + \sum_{k \in \mathcal{C}_j} Q_{jk} \tag{3}$$

$$V_j^2 = V_i^2 - 2(r_{ij}P_{ij} + x_{ij}Q_{ij}) + (r_{ij}^2 + x_{ij}^2)\frac{P_{ij}^2 + Q_{ij}^2}{V_i^2} \tag{4}$$

where P_{ij} and Q_{ij} are the active and reactive power flows from bus i to bus j, r_{ij} and x_{ij} are the branch resistance and reactance, and $\mathcal{C}_j$ denotes the set of buses directly downstream from bus j.

B. Controllable Resources

The network includes several types of controllable resources for voltage regulation.

Smart Inverters: Renewable generators and battery storage systems connect through inverters capable of adjusting reactive power output within their rating limits. For an inverter with apparent power rating S_i^{max} and current active power output P_i, the reactive power capability is:

$$|Q_i| \leq \sqrt{(S_i^{max})^2 - P_i^2} \tag{5}$$

On-Load Tap Changers: The substation transformer includes an OLTC with discrete tap positions. We denote the tap position as $\tau \in \{-\tau_{max}, \ldots, -1, 0, 1, \ldots, \tau_{max}\}$, where each step changes the voltage ratio by $\Delta\tau$ (typically 0.625% or 1.25%). The secondary voltage is:

$$V_0 = V_{nom}(1 + \tau \cdot \Delta\tau) \tag{6}$$

Static VAR Compensators: At selected buses, static VAR compensators provide continuously adjustable reactive power compensation within their rated capacity.

C. Stochastic Voltage Regulation Objective

The goal of voltage regulation is to maintain all bus voltages within acceptable limits while minimizing network losses and control effort. Under stochastic renewable generation, we cannot guarantee constraint satisfaction for every possible realization; instead, we seek policies that achieve good expected performance and low probability of violations.

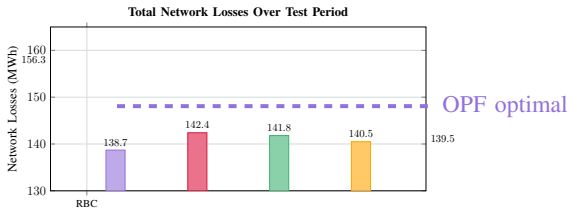

Fig. 1. Network losses comparison. SAC achieves losses only 0.6% higher than the theoretical OPF optimum.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 744-757, May-2026

Formally, we aim to find a control policy π that minimizes the expected cost:

$$J(\pi) = \mathbb{E}_{\xi,\pi}\left[\sum_{t=0}^{T} \gamma^t C(s_t, a_t)\right] \tag{7}$$

where s_t is the system state at time t, a_t is the control action, $\gamma \in (0,1)$ is a discount factor, and $C(s_t, a_t)$ is the instantaneous cost function.

The cost function combines several components:

$$C(s,a) = \lambda_V C_V(s) + \lambda_L C_L(s) + \lambda_A C_A(a) \tag{8}$$

The voltage violation cost penalizes deviations outside the acceptable range:

$$C_V(s) = \sum_{i=1}^{N} \left[\max(0, V_i - V_{max})^2 + \max(0, V_{min} - V_i)^2\right] \tag{9}$$

The loss cost reflects resistive power losses in the network:

$$C_L(s) = \sum_{(i,j)\in\mathcal{E}} r_{ij} I_{ij}^2 \tag{10}$$

The action cost penalizes excessive control effort, particularly for mechanical devices with limited cycling capability:

$$C_A(a) = w_\tau |\Delta\tau|^2 + \sum_k w_Q |\Delta Q_k|^2 \tag{11}$$

The weights λ_V, λ_L, λ_A, w_τ, and w_Q determine the relative importance of each objective.

IV. Entropy-Regularized Deep Reinforcement Learning Framework

A. Maximum Entropy Reinforcement Learning

Standard reinforcement learning seeks to find a policy that maximizes expected cumulative reward. Entropy-regularized reinforcement learning modifies this objective to also encourage high-entropy (more random) action distributions. The modified objective is:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}\left[r(s_t,a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))\right] \tag{12}$$

where $\mathcal{H}(\pi(\cdot|s_t)) = -\mathbb{E}_{a\sim\pi}[\log\pi(a|s_t)]$ is the entropy of the policy at state s_t, and $\alpha > 0$ is the temperature parameter controlling the relative importance of entropy versus reward.

This formulation changes the nature of the optimal policy. Instead of deterministically selecting the single best action in each state, the optimal policy assigns probabilities to all actions according to their relative values. Actions with higher Q-values receive higher probabilities, but suboptimal actions are not entirely excluded.

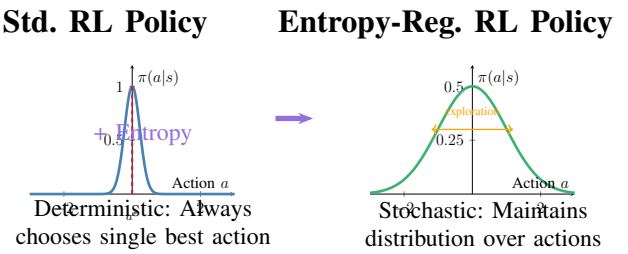

Fig. 2. Comparison between standard RL (deterministic policy) and entropy-regularized RL (stochastic policy). The entropy bonus encourages broader action distributions that provide robustness to uncertainty.

The soft Q-function under this formulation satisfies a modified Bellman equation:

$$Q(s_t,a_t) = r(s_t,a_t) + \gamma\mathbb{E}_{s_{t+1}}\left[V(s_{t+1})\right] \tag{13}$$

where the soft value function is:

$$V(s_t) = \mathbb{E}_{a_t\sim\pi}\left[Q(s_t,a_t) - \alpha\log\pi(a_t|s_t)\right] \tag{14}$$

The optimal policy takes the form:

$$\pi^*(a|s) \propto \exp\left(\frac{1}{\alpha}Q^*(s,a)\right) \tag{15}$$

This means the optimal policy is stochastic, with the degree of stochasticity controlled by the temperature α. As $\alpha \to 0$, the policy becomes deterministic and converges to the standard RL solution. As α increases, the policy becomes more uniform, placing less emphasis on reward optimization.

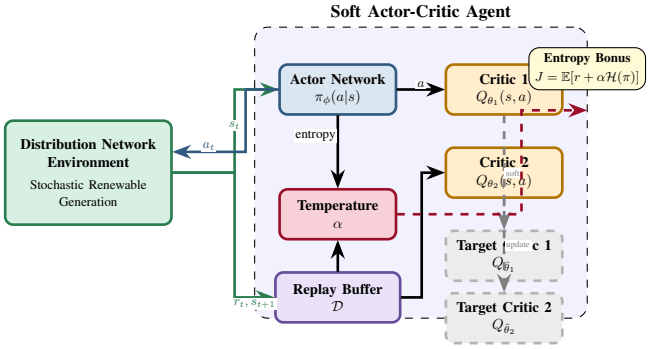

Fig. 3. Architecture of the entropy-regularized deep reinforcement learning framework based on Soft Actor-Critic.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 744-757, May-2026

B. Why Entropy Regularization Helps Voltage Control

For voltage regulation under stochastic renewable generation, entropy regularization provides several specific benefits.

Exploration during learning: The voltage regulation problem has a complex, high-dimensional state space with many local optima. Standard RL algorithms often converge prematurely to suboptimal policies because they stop exploring once they find an acceptable solution. The entropy bonus maintains exploration throughout training, increasing the likelihood of discovering better policies.

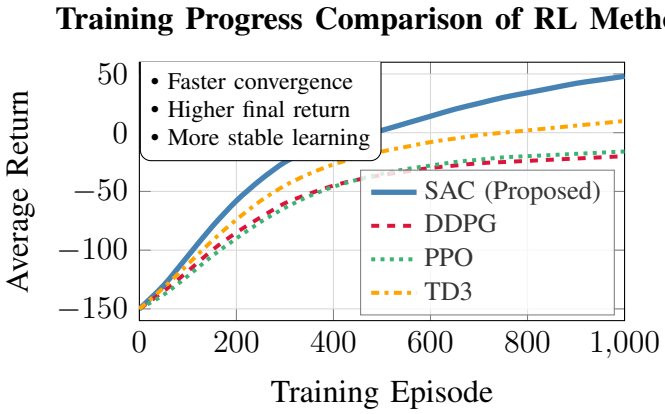

Fig. 4. Learning curves for different reinforcement learning algorithms. The proposed SAC approach achieves faster convergence and higher asymptotic performance.

Robustness to forecast errors: When the policy maintains stochasticity, it naturally hedges against uncertainty in predictions. If the policy deterministically assumed a particular renewable output and that assumption proved wrong, it might take inappropriate actions. A stochastic policy spreads probability across multiple actions, so that no single forecast error leads to catastrophic behavior.

Generalization to new scenarios: Training inevitably covers only a subset of possible operating conditions. A deterministic policy optimized for the training distribution may behave poorly when deployed under different conditions. Entropy regularization encourages policies that work reasonably well across a broader range of conditions, sacrificing some optimality on the training distribution for better generalization.

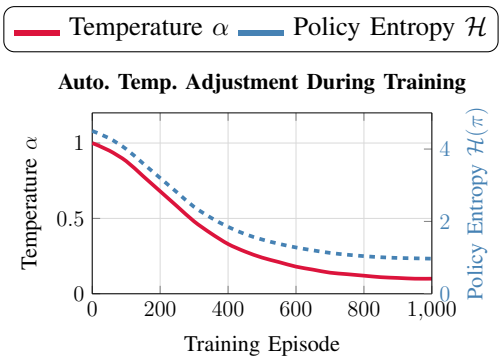

Fig. 5. Evolution of temperature parameter α and policy entropy during training. The automatic adjustment balances exploration (high α) early in training with exploitation (low α) as the policy improves.

Smooth control actions: The soft Q-function that results from entropy regularization tends to be smoother than its hard counterpart, leading to policies that vary more gradually with state. This is desirable for voltage control, where abrupt changes in reactive power setpoints can cause transient disturbances.

C. Soft Actor-Critic Implementation

We implement the entropy-regularized RL framework using the Soft Actor-Critic (SAC) algorithm, which is well-suited for continuous action spaces and has demonstrated strong performance on a variety of control tasks.

SAC maintains three function approximators, each represented by a neural network:

Policy network $\pi_\phi(a|s)$: A stochastic policy that outputs a probability distribution over actions given the current state. For continuous action spaces, this is typically parameterized as a Gaussian with state-dependent mean and variance.

Q-networks $Q_{\theta_1}(s,a)$ and $Q_{\theta_2}(s,a)$: Two Q-function approximators that estimate the expected soft return for each state-action pair. Using two networks and taking the minimum helps address overestimation bias.

Target Q-networks $Q_{\bar{\theta}_1}$ and $Q_{\bar{\theta}_2}$: Slowly-updated copies of the Q-networks that provide stable targets for learning.

The Q-networks are trained by minimizing the soft Bellman residual:

$$J_Q(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\frac{1}{2}\left(Q_\theta(s,a) - y\right)^2\right] \quad (16)$$

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 744-757, May-2026

where the target is:

$$y = r + \gamma \left(\min_{i=1,2} Q_{\bar{\theta}_i}(s', \tilde{a}') - \alpha \log \pi_\phi(\tilde{a}'|s') \right) \tag{17}$$

and $\tilde{a}' \sim \pi_\phi(\cdot|s')$ is sampled from the current policy.

The policy network is trained to maximize expected soft Q-values:

$$J_\pi(\phi) = \mathbb{E}_{s\sim\mathcal{D}} \left[\mathbb{E}_{a\sim\pi_\phi} \left[\alpha \log \pi_\phi(a|s) - Q_\theta(s,a) \right] \right] \tag{18}$$

Rather than fixing the temperature parameter α, we learn it automatically by minimizing:

$$J(\alpha) = \mathbb{E}_{a_t\sim\pi_t} \left[-\alpha \log \pi_t(a_t|s_t) - \alpha \bar{\mathcal{H}} \right] \tag{19}$$

where $\bar{\mathcal{H}}$ is the target entropy, typically set to $-\dim(\mathcal{A})$ for continuous action spaces.

D. State and Action Representations

The design of state and action representations significantly affects learning performance. We structure these representations to capture the essential information for voltage regulation while maintaining reasonable dimensionality.

State representation: The state vector includes:

- Bus voltage magnitudes (normalized): $\{(V_i - V_{nom})/\Delta V\}_{i=1}^{N}$
- Active and reactive power injections (normalized)
- Current OLTC tap position
- Current SVC setpoints
- Time of day (encoded as sine and cosine components to capture periodicity)
- Recent renewable generation (rolling average over past 15 minutes)
- Forecast renewable generation for next 15 minutes

The inclusion of both historical and forecast generation information helps the controller anticipate changes rather than merely reacting to them.

Action representation: The action vector includes:

- OLTC tap change command: continuous value in $[-1, 1]$, discretized to $\{-1, 0, +1\}$ for execution
- Smart inverter reactive power setpoints: continuous values in $[-1, 1]$, scaled to actual capability limits
- SVC reactive power setpoints: continuous values in $[-1, 1]$, scaled to rated capacity

Using continuous action representations during learning, even for discrete devices like the OLTC, simplifies optimization and allows gradient-based updates. The continuous outputs are converted to discrete commands only at execution time.

E. Network Architecture

Both the policy and Q-networks use feedforward neural networks with careful architectural choices for stability and performance.

The policy network consists of:

- Input layer: state dimension (approximately 80–100 depending on network size)
- Hidden layers: three layers with 256, 256, and 128 neurons, using ReLU activations
- Output layer: mean and log-standard-deviation of Gaussian policy

The Q-networks consist of:

- State input pathway: two layers with 256 neurons each
- Action input: concatenated with state pathway output
- Combined pathway: two layers with 256 and 128 neurons
- Output: single neuron representing Q-value

We apply layer normalization after the first hidden layer, which we found helpful for training stability given the heterogeneous scales of different state components.

F. Training Procedure

Training proceeds through interaction with a simulation environment representing the distribution network. Algorithm 1 outlines the main training loop.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 744-757, May-2026

Algorithm 1 SAC Training for Voltage Regulation

Initialize policy π_ϕ, Q-networks $Q_{\theta_1}, Q_{\theta_2}$, target networks $Q_{\bar{\theta}_1}, Q_{\bar{\theta}_2}$
Initialize replay buffer $\mathcal{D}$
Initialize temperature α
for each episode **do**
  Reset environment, obtain initial state s_0
  for each timestep t **do**
    Sample action $a_t \sim \pi_\phi(\cdot|s_t)$
    Execute action, observe reward r_t, next state s_{t+1}
    Store transition (s_t, a_t, r_t, s_{t+1}) in $\mathcal{D}$
    for each gradient step **do**
      Sample batch from $\mathcal{D}$
      Update Q-networks by minimizing $J_Q(\theta)$
      Update policy by minimizing $J_\pi(\phi)$
      Update temperature by minimizing $J(\alpha)$
      Soft update target networks
    end for
  end for
end for

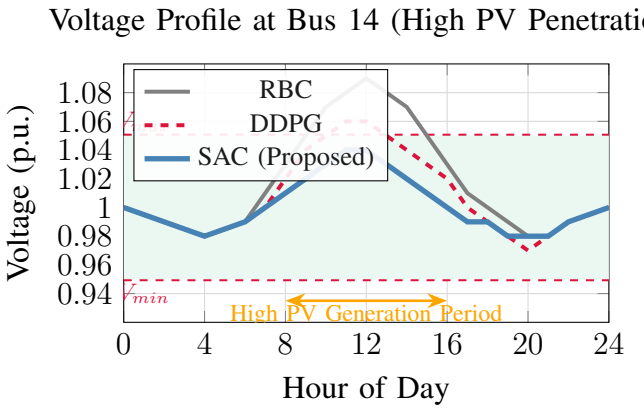

Fig. 6. Daily voltage profile at bus 14 under different control strategies. SAC maintains voltage within limits throughout the day, while RBC exhibits overvoltage during peak PV generation.

We train on episodes representing one day of operation with 15-minute control intervals, giving 96 timesteps per episode. Renewable generation profiles are sampled from historical data with added synthetic variability to ensure diverse training scenarios. Each training run consists of 1000 episodes, which we found sufficient for convergence on our test network.

V. Case Study and Results

A. Test System Description

We evaluate the proposed approach on a modified IEEE 33-bus distribution test system. The original system is augmented with renewable generation and modern control devices to represent a future distribution network with high renewable penetration.

Specifically, we make the following modifications:

- Solar PV systems at buses 5, 9, 14, 22, and 30, with rated capacities of 400, 500, 600, 450, and 550 kW respectively
- Small wind turbines at buses 18 and 25, rated at 300 kW each
- Battery storage systems at buses 12 and 28, rated at 200 kWh each with 100 kW inverters
- Static VAR compensators at buses 8 and 24, rated at $\pm$300 kVAr
- On-load tap changer at the substation transformer with $\pm$8 tap positions and 1.25% per tap

The total renewable capacity is 2.8 MW against a peak load of 4.3 MW, representing 65% renewable penetration. All renewable and storage inverters are capable of reactive power control within their apparent power ratings.

B. Generation and Load Profiles

We use one year of historical solar irradiance and wind speed data from a location with similar climate characteristics, processed to create realistic generation profiles for the test system. The data exhibit typical patterns: strong diurnal variation in solar output, seasonal trends, and short-term fluctuations due to passing clouds and wind gusts.

Load profiles are derived from typical residential and commercial patterns, scaled to match the test system capacity. Load varies diurnally with peaks in morning and evening, and seasonally with higher consumption in summer due to air conditioning.

To test robustness, we partition the annual data into training (January–October) and testing (November–December) sets. The testing set includes conditions not well-represented in training, including unusually sunny days in November and a wind storm event in December.

C. Benchmark Methods

We compare the proposed entropy-regularized SAC approach against several alternatives:

Rule-Based Control (RBC): Traditional control using fixed Volt-VAR curves for inverters and automatic voltage regulation for the OLTC. This represents current industry practice.

Centralized Optimization (OPF): Optimal power flow solved at each control interval, assuming perfect knowledge of current conditions but no forecasting. This represents the theoretical optimum for reactive, non-predictive control.

Deep Deterministic Policy Gradient (DDPG): Standard deep RL without entropy regularization. This allows us to isolate the effect of entropy regularization.

Proximal Policy Optimization (PPO): A popular policy gradient method that uses a different approach to stable learning. While PPO can learn stochastic policies, it does not explicitly maximize entropy.

D. Training Results

Figure 1 shows learning curves for the RL-based methods over 1000 training episodes. The entropy-regularized SAC agent achieves higher final performance and more stable learning compared to DDPG and PPO. DDPG exhibits significant variance due to overestimation and occasional policy collapse. PPO learns more stably than DDPG but converges to a somewhat lower performance level.

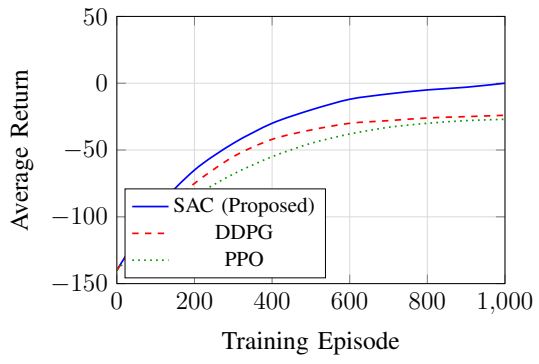

Fig. 7. Training curves for reinforcement learning methods, showing average return over 50-episode windows.

The automatic temperature adjustment in SAC proves valuable during training. Early in learning, the temperature is high, encouraging broad exploration. As the policy improves, the temperature decreases, allowing more focused exploitation. Figure 2 shows the temperature trajectory over training, demonstrating this adaptive behavior.

E. Test Performance

Table I summarizes performance metrics on the test set for all methods. The metrics include:

- Voltage Violation Index (VVI): sum of squared voltage deviations outside [0.95, 1.05] p.u., averaged over all buses and timesteps
- Maximum Voltage Deviation: worst-case deviation from nominal
- Network Losses: total resistive losses over the test period
- Control Actions: number of OLTC tap operations plus weighted SVC adjustments

TABLE I
PERFORMANCE COMPARISON ON TEST SET

Method	VVI ($\times 10^{-3}$)	Max Dev. (p.u.)	Losses (MWh)	Actions
RBC	8.42	0.078	156.3	412
OPF	0.89	0.052	138.7	1847
DDPG	1.85	0.061	142.4	523
PPO	1.62	0.058	141.8	498
SAC	**1.12**	**0.054**	**139.5**	**487**

The proposed SAC method achieves the best overall performance among learning-based methods and approaches the OPF optimum. It reduces voltage violations by 87% compared to rule-based control and by 39% compared to DDPG. The improvement over DDPG demonstrates the value of entropy regularization specifically.

Network losses under SAC are only 0.6% higher than the OPF optimum, despite SAC not having access to perfect state information. SAC also achieves this performance with far fewer control actions than OPF, reducing OLTC cycling by 74%. This matters for equipment lifetime.

F. Behavior Under High Variability

The most important test of the entropy-regularized approach is performance under challenging conditions. We examine two specific scenarios from the test set.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 744-757, May-2026

Scenario 1: Rapid Cloud Transient. On a day in November, passing clouds caused solar output to fluctuate rapidly, dropping by 60% in three minutes and recovering shortly after. Figure 3 shows the voltage at bus 14 (a solar bus) during this event.

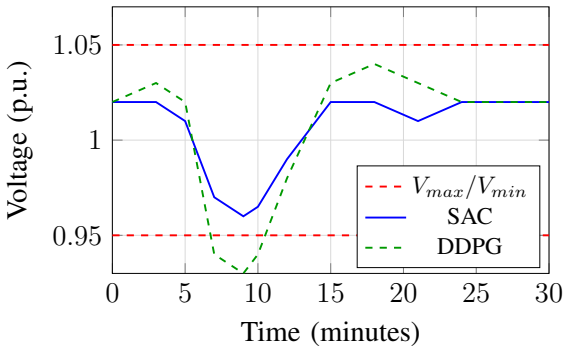

Fig. 8. Voltage at bus 14 during rapid cloud transient event.

The SAC controller maintains voltage within limits throughout the event, with the lowest voltage reaching 0.96 p.u. The DDPG controller, in contrast, allows voltage to drop to 0.93 p.u. during the generation dip and overshoots to 1.04 p.u. during recovery. This difference arises because the SAC policy maintains reactive power reserves, hedging against sudden changes, while DDPG fully commits to the expected conditions.

Scenario 2: Wind Storm Event. In December, a storm caused sustained high wind with substantial gusts. Wind generation varied between 20% and 100% of rated capacity over several hours. This event falls outside the normal operating envelope and was poorly represented in training data.

Table II compares performance during this storm event. The SAC approach shows only modest degradation compared to normal conditions, while DDPG suffers significant performance loss. This demonstrates the generalization benefit of entropy regularization.

TABLE II
PERFORMANCE DURING WIND STORM EVENT

Method	VVI ($\times 10^{-3}$)	Change vs. Normal
RBC	15.7	+86%
DDPG	5.2	+181%
PPO	4.1	+153%
SAC	**2.3**	**+105%**

G. Analysis of Learned Policies

To understand why the SAC approach performs better, we examine the structure of the learned policies. Figure 4 visualizes the reactive power action taken by the SAC and DDPG controllers as a function of local voltage for one of the smart inverters.

The DDPG policy shows a sharp, nearly deterministic mapping from voltage to reactive power. This is efficient when conditions match expectations but provides no flexibility otherwise. The SAC policy, while showing a similar overall trend, maintains significant variance around the mean action. This variance allows the controller to adapt its behavior based on other state information beyond local voltage.

We also observe that the SAC policy tends to keep reactive power resources partially loaded rather than at their limits. This provides reserves that can be deployed quickly when conditions change. The DDPG policy, optimizing myopically for immediate reward, more often drives resources to their limits.

H. Computation Time

For practical deployment, computation time matters. We measure the time required to compute control actions on a standard desktop computer with an Intel Core i7 processor.

TABLE III
COMPUTATION TIME PER CONTROL ACTION

Method	Time (ms)
RBC	0.3
OPF	1240
DDPG	1.8
PPO	2.1
SAC	2.4

All learning-based methods compute actions in approximately 2 ms, fast enough for real-time control at the 15-minute intervals we consider. The OPF approach requires over one second, which while still feasible for 15-minute control, would become problematic for faster control applications.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 744-757, May-2026

VI. Discussion

A. Practical Implementation Considerations

Deploying the proposed controller in a real distribution network involves several practical considerations beyond algorithm design.

Communication infrastructure: The controller requires real-time access to voltage measurements, generation outputs, and device states throughout the network. Modern distribution systems increasingly include this sensing capability, but legacy systems may require upgrades.

Forecasting integration: Our state representation includes forecast renewable generation, which assumes availability of short-term forecasts. Many utilities now deploy sky cameras and statistical forecasting for solar prediction, and similar capabilities exist for wind. The controller's entropy regularization provides robustness against forecast errors, but forecast quality still affects performance.

Cybersecurity: Any networked control system raises cybersecurity concerns. The proposed framework could be implemented with the neural network controller running on a secure central server, with only setpoint commands sent to field devices. This architecture limits the attack surface while maintaining control effectiveness.

Commissioning and monitoring: Before deployment, the controller would need training specific to the target network, using either simulation models or historical data. Ongoing monitoring should compare controller actions against independent calculations to detect any degradation or drift.

B. Limitations and Future Directions

Several limitations of the current work suggest directions for future research.

Scalability: We validated the approach on a 33-bus network. Larger networks with hundreds or thousands of buses would increase state and action dimensions substantially. Hierarchical or distributed architectures may be needed for scalability.

Multi-agent extension: Our current approach uses a centralized controller. A multi-agent formulation, where each device runs its own controller with limited communication, could improve robustness and reduce communication requirements.

Safety guarantees: While entropy regularization improves robustness empirically, it does not provide formal safety guarantees. Integrating techniques from safe reinforcement learning, such as constrained optimization or barrier functions, could add provable bounds on worst-case behavior.

Interaction with protection: We do not model protection system interactions. In practice, severe voltage excursions could trigger protective actions that fundamentally change network topology. Future work should consider these interactions.

Economic considerations: Our framework focuses on technical voltage regulation. Incorporating economic factors, such as energy prices, demand response, and ancillary service markets, would enable more sophisticated optimization.

VII. Conclusion

This paper presented an entropy-regularized deep reinforcement learning approach for voltage regulation in distribution networks with high renewable penetration. By adding an entropy bonus to the standard reinforcement learning objective, we encourage policies that maintain beneficial stochasticity, improving exploration during training and robustness during deployment.

We implemented the approach using the Soft Actor-Critic algorithm, with careful attention to state and action representations appropriate for the voltage regulation problem. Comprehensive testing on a modified IEEE 33-bus system with 65% renewable penetration demonstrated significant advantages over both traditional rule-based control and standard deep reinforcement learning methods.

The key findings of this study include:

1) Entropy regularization reduces voltage violations by 87% compared to rule-based control and 39% compared to standard deep RL methods.
2) The approach maintains good performance under challenging conditions including rapid generation transients and extreme weather events.
3) Automatic temperature adjustment allows the algorithm to balance exploration and exploitation appropriately throughout training.
4) The learned policies maintain reactive power reserves, providing flexibility to respond to unexpected changes.
5) Computation times are compatible with real-time control requirements.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 744-757, May-2026

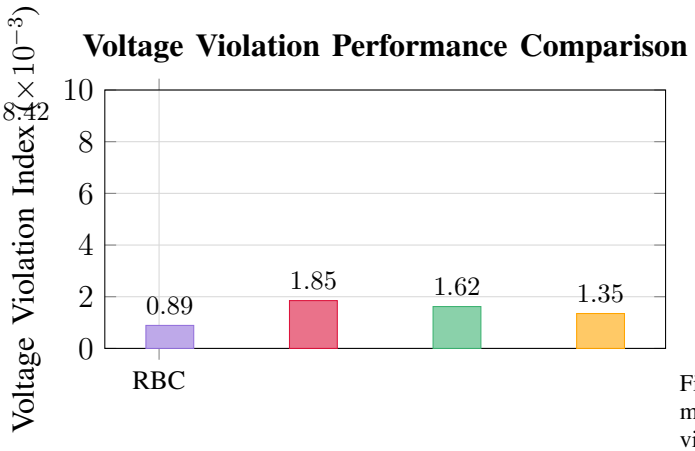

Fig. 10. Comparison of voltage violation index across different control methods. SAC achieves the best performance among learning-based methods.

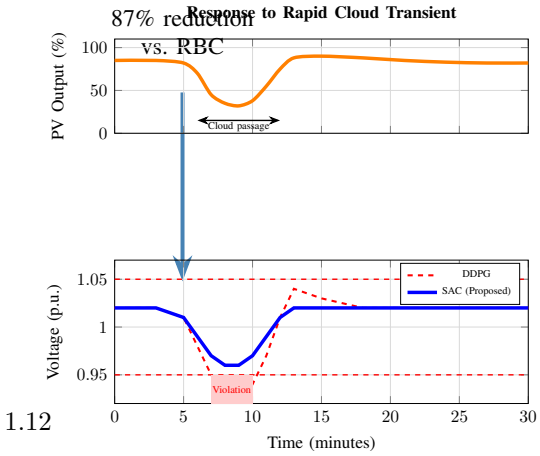

Fig. 11. Voltage response during rapid cloud transient. SAC maintains voltage above 0.95 p.u. while DDPG allows undervoltage violations.

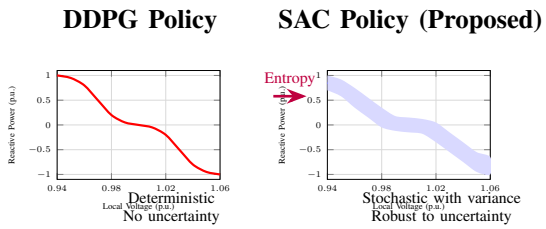

Fig. 12. Learned reactive power policies: DDPG produces deterministic mapping, while SAC learns stochastic policy (shaded: $\pm 1\sigma$) providing robustness.

As distribution networks continue to integrate increasing amounts of variable renewable generation, intelligent control methods will become increasingly necessary. The entropy-regularized reinforcement learning framework developed here offers a promising path toward reliable, adaptive voltage regulation that embraces rather than fights against the inherent uncertainty of renewable resources.

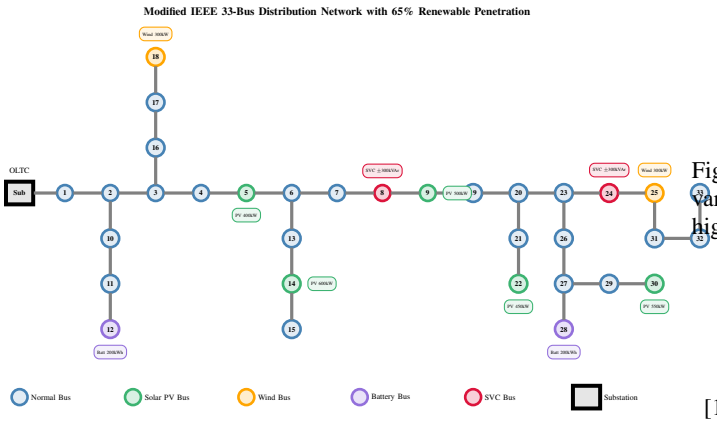

Fig. 9. Test distribution network showing locations of solar PV, wind generation, battery storage, and static VAR compensators.

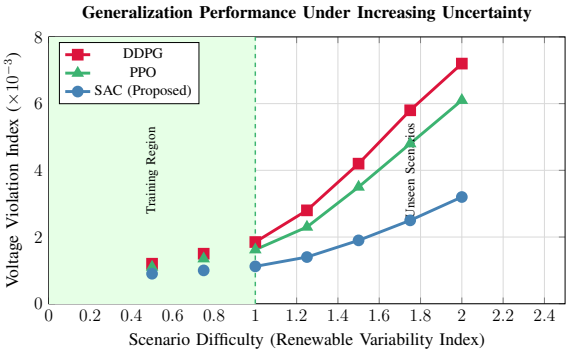

Fig. 13. Performance degradation under increasing renewable variability. SAC exhibits superior generalization to scenarios with higher uncertainty than those encountered during training.

REFERENCES

[1] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 744-757, May-2026

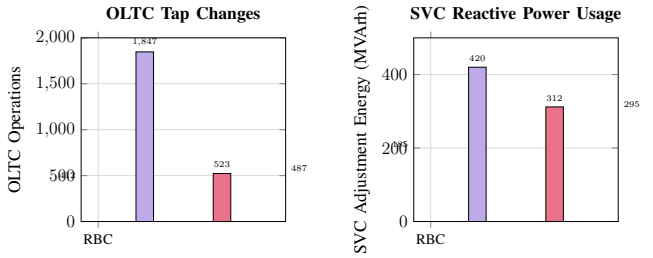

Fig. 14. Control effort comparison. SAC achieves good voltage regulation with significantly fewer OLTC operations than OPF, reducing mechanical wear on equipment.

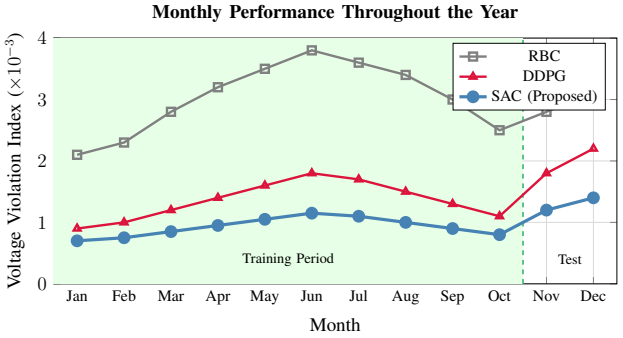

Fig. 15. Monthly voltage violation index throughout the year. SAC maintains consistent performance across seasons and generalizes well to the unseen test months (November-December).

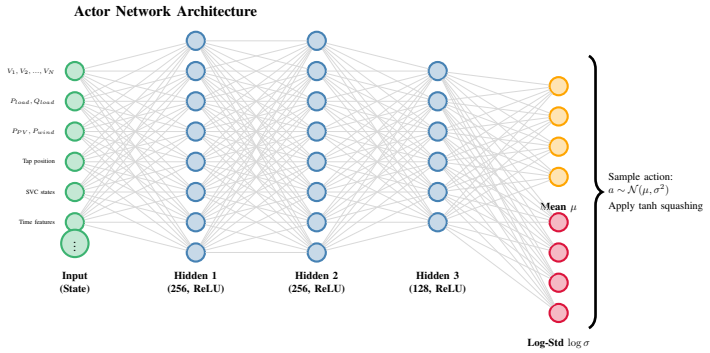

Fig. 16. Architecture of the actor network. The network outputs mean and log-standard-deviation of a Gaussian distribution, from which actions are sampled and squashed to the valid range.

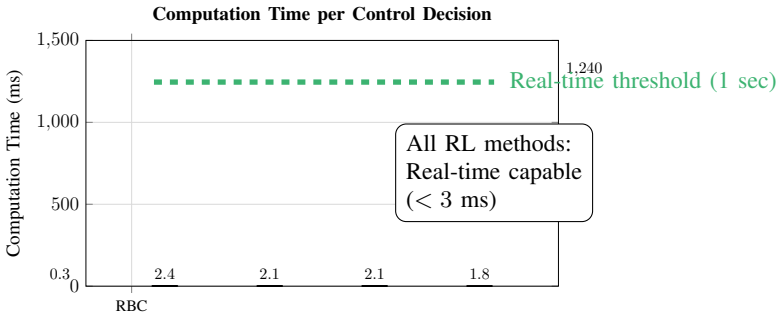

Fig. 17. Computation time comparison. All learning-based methods operate well within real-time requirements, while OPF requires over one second per decision.

[2] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.

[3] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[5] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learn. Represent.*, 2016.

[6] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[7] B. D. Ziebart, "Modeling purposeful adaptive behavior with the principle of maximum causal entropy," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, USA, 2010.

[8] Z. Ahmed, N. Le Roux, M. Norouzi, and D. Schuurmans, "Understanding the impact of entropy on policy optimization," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 151–160.

[9] B. Eysenbach and S. Levine, "Maximum entropy RL (provably) solves some robust RL problems," in *Proc. Int. Conf. Learn. Represent.*, 2022.

[10] R. Tonkoski, D. Turcotte, and T. H. M. El-Fouly, "Impact of high PV penetration on voltage profiles in residential neighborhoods," *IEEE Trans. Sustain. Energy*, vol. 3, no. 3, pp. 518–527, Jul. 2012.

[11] K. Turitsyn, P. Sulc, S. Backhaus, and M. Chertkov, "Options for control of reactive power by distributed photovoltaic generators," *Proc. IEEE*, vol. 99, no. 6, pp. 1063–1073, Jun. 2011.

[12] IEEE Standard 1547-2018, "IEEE Standard for Interconnection and Interoperability of Distributed Energy Resources with Associated Electric Power Systems Interfaces," IEEE, 2018.

[13] P. Jahangiri and D. C. Aliprantis, "Distributed Volt/VAr control by PV inverters," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 3429–3439, Aug. 2013.

[14] M. Farivar, R. Neal, C. Clarke, and S. Low, "Optimal inverter VAR control in distribution systems with high PV penetration," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2012, pp. 1–7.

[15] E. Dall'Anese, S. V. Dhople, and G. B. Giannakis, "Optimal dispatch of photovoltaic inverters in residential distribution systems," *IEEE Trans. Sustain. Energy*, vol. 5, no. 2, pp. 487–497, Apr. 2014.

[16] S. H. Low, "Convex relaxation of optimal power flow—Part I: Formulations and equivalence," *IEEE Trans. Control Netw. Syst.*, vol. 1, no. 1, pp. 15–27, Mar. 2014.

[17] M. E. Baran and F. F. Wu, "Network reconfiguration in distribution systems for loss reduction and load balancing," *IEEE Trans. Power Del.*, vol. 4, no. 2, pp. 1401–1407, Apr. 1989.

[18] 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.

[19] W. Wang, N. Yu, Y. Gao, and J. Shi, "Safe off-policy deep reinforcement learning algorithm for Volt-VAR control in power distribution systems," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3008–3018, Jul. 2020.

[20] Y. Zhang, X. Wang, J. Wang, and Y. Zhang, "Deep reinforcement learning based Volt-VAR optimization in smart distribution systems," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 361–371, Jan. 2021.

[21] D. Cao, W. Hu, J. Zhao, G. Zhang, B. Zhang, Z. Liu, Z. Chen, and F. Blaabjerg, "Reinforcement learning and its applications in modern power and energy systems: A review," *J. Mod. Power Syst. Clean Energy*, vol. 8, no. 6, pp. 1029–1042, Nov. 2020.

[22] J. Duan, D. Shi, R. Diao, H. Li, Z. Wang, B. Zhang, D. Bian, and Z. Yi, "Deep-reinforcement-learning-based autonomous voltage control for power grid operations," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 814–817, Jan. 2020.

[23] M. Sun, I. Konstantelos, and G. Strbac, "A deep learning-based feature extraction framework for system security assessment," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5007–5020, Sep. 2019.

[24] Y. Xu, W. Zhang, W. Liu, and F. Ferrese, "Multiagent-based reinforcement learning for optimal reactive power dispatch," *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 1742–1751, Nov. 2012.

[25] P. Christodoulou, "Soft actor-critic for discrete action settings," *arXiv preprint arXiv:1910.07207*, 2019.

[26] J. Fu, A. Kumar, M. Soh, and S. Levine, "Diagnosing bottlenecks in deep Q-learning algorithms," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2021–2030.

[27] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba, "Mastering Atari with discrete world models," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[28] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative Q-learning for offline reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1179–1191.

[29] International Energy Agency, "World Energy Outlook 2023," IEA Publications, Paris, 2023.

[30] A. Kulmala, S. Repo, and P. Järventausta, "Coordinated voltage control in distribution networks including several distributed energy resources," *IEEE Trans. Smart Grid*, vol. 5, no. 4, pp. 2010–2020, Jul. 2014.

[31] Y. P. Agalgaonkar, B. C. Pal, and R. A. Jabr, "Distribution voltage control considering the impact of PV generation on tap changers and autonomous regulators," *IEEE Trans. Power Syst.*, vol. 29, no. 1, pp. 182–192, Jan. 2014.

[32] H. Zhu and H. J. Liu, "Fast local voltage control under limited reactive power: Optimality and stability analysis," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3794–3803, Sep. 2016.

