# Offline Reinforcement Learning Agents for Adaptive Reactive Power Control with Renewable Energy Sources

Tejashri Bhatt, Stephan Balduin, and Eric MSP Veith
Institute for Information Technology
OFFIS e.V.
Oldenburg, Germany
e-mail: {tejashri.bhatt | stephan.balduin | eric.veith}@offis.de

***Abstract*—Conventional reactive power control is typically performed by operators through coordinated switching of power electronic devices. This task is becoming increasingly complex as the integration of renewable energy sources, such as rooftop photovoltaic systems and wind turbines, expands. Maintaining grid stability is critical to ensure energy supply without risking equipment damage. In this context, artificial Reinforcement Learning (RL) agents for reactive power control can assist operators by suggesting actions, though final decisions remain with the operator. High-performing automated RL algorithms are essential for this as they enable execution of complex actions through trial and error, facilitating the adaptable transfer of learning to the real world. While established algorithms, such as Soft Actor-Critic (SAC), Deep Deterministic Policy Gradient (DDPG), Twin-Delayed DDPG and Proximal Policy Optimization (PPO), offer solutions, each has limitations. Training artificial RL agents in real-world power grids is impractical due to the safety-critical concerns, stressing the need for an alternative approach. SAC provides benefits in continuous action space, such as improved exploration and leveraging past experiences, but suffers from long training times. This paper addresses the issue by reducing SAC training periods through the integration of the Behavior Cloning from Observation (BCO) algorithm. This approach enhances performance by initializing SAC with a high-performing, pre-trained Artificial Neural Network (ANN) rather than a random policy, providing a strong starting point while preserving the benefits of SAC.**

***Keywords*-Smart Grid Management; Reactive Power Control; Artificial Intelligence; Soft Actor-Critic; Behavioral Cloning from Observation; Renewable Energy Integration; Offline Reinforcement Learning.**

## I. Introduction

Autonomous systems hold significant potential for power systems, a domain where mismanagement can have extensive societal repercussions. As power systems evolve with increasing dynamic complexity and renewable integration, traditional control methods are becoming inadequate. Autonomous systems provide real-time decision-making, optimize resource allocation and adapt to fluctuating conditions with minimal human input. This transition to autonomy not only enhances operational efficiency but also ensures grid reliability amid rapid technological advancements and energy demands.

Reinforcement Learning (RL), as defined by Sutton et al. [1], is *"learning what to do—how to map situations to actions—so as to maximize a numerical reward signal."* Here, an agent learns by interacting with its environment, receiving feedback in the form of rewards to guide its actions. Two key aspects of RL are its reliance on trial-and-error learning and its consideration of delayed rewards.

Offline RL, or batch RL, trains agents on static datasets, which is advantageous in fields like healthcare, autonomous driving and power systems, where real-time data gathering can be costly or risky [2]. Unlike online RL, which continuously interacts with the environment, offline RL relies on pre-existing datasets to learn policies, minimizing costs and enhancing safety. However, a challenge in offline learning is concerning distributional shift, which arises when the training data differs significantly from real-world scenarios, necessitating high-quality datasets to ensure reliable outcomes.

Behavior Cloning from Observation (BCO) is a notable offline RL algorithm that enables an agent to learn tasks by observations only, without direct access to an expert's actions. By leveraging pre-existing data, BCO mimics expert behavior and iteratively refines policies to perform accurately in complex settings. If the data is optimal, BCO can help overcome common challenges in offline RL, providing robust, efficient learning strategies.

Distributed generators, like Photovoltaic (PV) systems with inverters, offer strategic advantages for reactive power control at key locations. Effective voltage control and Reactive Power Management (RPM) are crucial for power system reliability, typically managed through centralized control by system operators [3]. These operators use comprehensive system data and advanced computer models to make informed decisions, as suppliers do not have direct control over voltage control needs. The operators would benefit from quick assistance in making timely decisions, with a potential Artificial Intelligence (AI) agent suggesting corrective actions while leaving the final decision to the operator. This seamless integration of centralized control and distributed resources could enhance the efficiency and reliability of voltage stability and RPM in power systems.

The primary contributions of this paper are as follows:

1) Addressing controller conflicts arising from physical constraints, such as the impossibility of achieving a 1 voltage magnitude per unit (p.u.) across multiple buses in series simultaneously. In standard Soft Actor-Critic (SAC), an exploration dilemma emerges, as the agent incorrectly assumes actuator independence across buses, where, in reality, actuator behavior is interdependent due to the physical limitations. To address this, we developed a specialized RL agent tailored for this RPM context.



 ISBN: 978-1-68558-242-5 

2) Applying BCO to expedite training while maintaining key benefits of off-policy learning and robust exploration. The use of BCO directly aids in overcoming the exploration challenge outlined above, where agents benefit from a structured observational learning approach.
3) Demonstrating the scalability and transferability of the proposed approach for application to more complex tasks within the power systems domain.

This study is constrained by the following factors:

1) The PalaestrAI framework is used for reactive power control implementation.
2) SAC is selected for policy learning due to its suitability for continuous action spaces and superior stability over Deep Deterministic Policy Gradient (DDPG) and Proximal Policy Optimization (PPO) [4].
3) BCO is preferred due to the availability of high-quality MIDAS data (discussed later), making Advantage Weighted Actor-Critic (AWAC) unnecessary [5].
4) Each one-year simulation requires one hour, with additional time for training and testing.

The remainder of this paper is structured as follows: Section II surveys the related work, covering key literature on SAC, BCO and recent advancements in RPM utilizing AI, in addition to open-source tools used in this work such as Midas and palaestrAI. Section III presents the methodology, describing the grid environment and grid code utilized. It also covers the development of three scenarios, experiment setup and the performance metrics applied for evaluation. Section IV presents the results while Section V provides an analysis and discussion of the results for each scenario, as well as ablation experiment. Lastly, Section VI summarizes the main findings and offers potential avenues for future research.

## II. Related Work

This section examines the integration of SAC, a state-of-the-art RL method, with BCO, for reactive power control, leveraging historical data and entropy-based learning for enhanced stability. It also reviews advancements in AI-driven reactive power management, identifies research gaps and highlights the open-source tools Midas and palaestrAI for scalable and reproducible simulations.

### A. Soft Actor-Critic

The SAC algorithm, introduced by Haarnoja et al. [6], is an off-policy maximum entropy actor-critic framework designed to balance exploration with reward maximization. In entropy-regularized reinforcement learning, entropy—representing randomness in a policy—adds a bonus reward at each step. This motivates agents to explore more by maximizing both the cumulative reward and entropy [7] [8]. This entropy-enhanced approach provides sample-efficient learning, stability and adaptability to complex tasks. For further details on SAC and its implementation, refer to [4] and [7]. The resulting objective function is:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\left(R(s_t,a_t,s_{t+1}) + \alpha H\left(\pi(\cdot|s_t)\right)\right)\right] \tag{1}$$

where $\alpha > 0$ is a regularization factor. The modified value functions $V_\pi(s)$ and $Q_\pi(s,a)$ now include entropy terms, with the following transformations:

$$V_\pi(s) = \mathbb{E}_{a\sim\pi}[Q_\pi(s,a)] + \alpha H\left(\pi(\cdot|s)\right) \tag{2}$$

$$Q_\pi(s,a) = \mathbb{E}_{s'\sim P, a'\sim\pi}[R(s,a,s') + \gamma(V_\pi(s') + \alpha H(\pi(\cdot|s')))] \tag{3}$$

This entropy-enhanced approach provides sample-efficient learning, stability and adaptability to complex tasks. For further details on SAC and its implementation, refer to [4] and [7].

### B. Behavior Cloning from Observation

As outlined in the introduction, BCO is a learning approach that enables an agent to mimic expert actions by observing state transitions, bypassing the need for explicit action data and allowing skill acquisition without complete state-action mappings.

The BCO framework begins by initializing policy and model training with an offline dataset, eliminating the need for real-time environment interaction during this stage. Once an initial policy is derived, it is refined through online RL. A common challenge in BCO is the risk of distributional shift, where changes in the offline dataset adversely impact performance, as noted by Prudencio et al. [2]. Thus, a reliable dataset, ideally from expert sources, is crucial for initial training.

In this study, BCO is combined with SAC, which uses entropy-based learning, enhancing the agent's capacity for complex decision-making tasks. Together, BCO and SAC principles contribute to accelerated learning and increased adaptability in reinforcement learning settings. For a more detailed BCO understanding, refer to [9].

### C. Progress in Reactive Power Management via AI

Prior research has focused on using artificial intelligence with renewable energy sources to enhance reactive power control and voltage stability in smart grids. For instance, Chandrasekaran et al. [10] propose a hybrid model using solar and wind energy with Artificial Neural Network (ANN) and Distribution Static Synchronous Compensator (DSTATCOM), achieving a voltage profile accuracy of $98.45\,\%$ and reducing real power loss by $15\,\%$. Similarly, Utama et al. [11] leverage ANN-based controllers to manage reactive power in the CIGRE Medium Voltage (MV) grid with PV systems, addressing issues such as voltage fluctuations and line congestion in both centralized and decentralized frameworks. To improve smart grid operations, Fiorotti et al. [12] apply a Genetic Algorithm for optimal active and reactive power management, decreasing

net present value by 28.15 % and energy costs by 78.16 % by adapting to diverse consumption profiles. However, an approach is needed that can better adapt to complex real-world scenarios and leverage the wealth of historical data effectively. In this context various RL algorithms can be highly advantageous.

Rehman et al. [13] investigate reactive power control in PV inverters through a decentralized actor-critic RL framework, achieving stable voltage control ratios and minimized power loss, with voltage levels consistently within 0.95-1.05 p.u. The authors suggest exploring alternative algorithms for enhanced control in future work. Wolgast et al. [14] employed RL agents for voltage control, with a focus on the impact of environment definitions on performance; however, their study did not explore the influence of advanced RL algorithms. Addressing these limitations, this paper introduces a more advanced algorithm, SAC, over the actor-critic approach and incorporates an ANN for initialization instead of random policy initialization, thereby also implementing BCO.

### D. Research Gap

Applying BCO through SAC in RL benchmark environments has, to date, been explored only in one master's thesis by D'Silva et al. [15]. To our knowledge, the application of BCO with SAC specifically in reactive power control for MV grids remains unexplored. Prior research by Dey et al. [16] has utilized BCO with PPO for building energy control; however, the on-policy nature of PPO limits its efficiency in multi-task settings. By substituting PPO with SAC, an off-policy algorithm, this study overcomes these constraints, enabling more efficient training and enhanced grid management performance. This approach extends BCO to grid integration scenarios, leveraging abundant historical data and an ANN-trained policy to initialize SAC for managing reactive power control in an MV grid.

This research investigates whether constructing BCO by integrating the SAC algorithm with ANN policy initialization can reduce training latency for SAC agents in achieving effective reactive power control.

### E. Open-Source Tools

*1) Midas:* Midas is an open-source framework that allows easy configuration of a power system co-simulation [17] and uses the co-simulation framework Mosaik as backend [18]. It features various time series for consumers and producers, weather time series and simulation models of renewable energy sources like photovoltaic, wind and biogas, as well as a battery and a cold warehouse. Midas is configured with scenario files in YAML format, which specify what kind of load or generation is connected to certain buses in the grid. The framework provides a seamless integration into palaestrAI, which is described in the following section.

*2) PalaestrAI:* The palaestrAI framework is instrumental in simulating real-world scenarios for electricity grid integration, offering a comprehensive set of components. It is designed to implement the Adversarial Resilience Learning (ARL) methodology by Veith et al. [19]. It encompasses various packages, emphasizing a reliable experimentation process through experiment definitions and proper data storage. Users can easily create experiment files to achieve reproducible simulations, defining environments, agents and their parameters for experimentation within palaestrAI.

### F. Objective Function

The reward function of the RL agent used for this study is composed of four main components, each associated with specific weighting variables represented by the world state of the system in terms of voltage levels across all buses, buses particularly under control, the bus status based on the impact of grid code violations and the quantity of real power production within the grid relative to total demand. For more details on the formulation of objective function please refer [20].

## III. Methodology

All experiments in this paper utilize the palaestrAI framework from Veith et al. [19] focusing on a single agent. Each experiment includes distinct training and testing phases, with a maximum simulated duration of one year and a 15-minute interval for each step. The overall process flow for methodology is depicted in Figure 2.

### A. Grid Environment

In this study, a 20 kV MV grid is connected to a 110 kV transmission network, with a total power capacity of 2000 kW. The grid is modeled after the CIGRE MV benchmark grid [21], comprising 14 buses, each equipped with a PV generator with randomly assigned output for variability. Weather data from Bremen, Germany (see Figure 1) and static load time series from the Midas project simulate realistic conditions [17]. Additionally, commercial loads, such as a supermarket and a small hotel, enrich the grid's complexity.

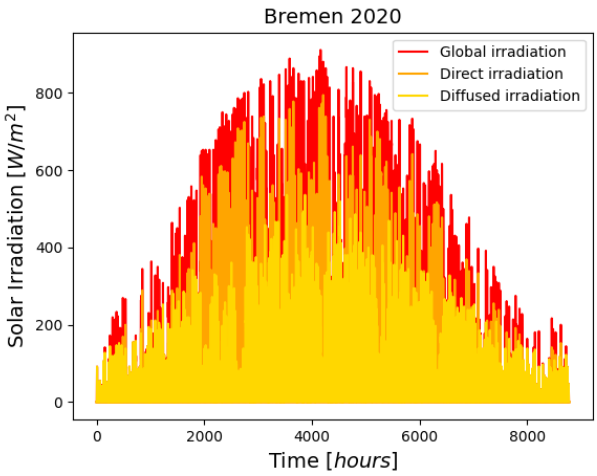



Figure 1. Global Solar Radiation in Bremen, Germany 2020.

This paper follows technical guidelines (VDE-AR-N 4110) published by the German Association for Electrical, Electronic & Information Technologies for MV grids in Germany,

ensuring compliance with standards for the safe integration of renewable energy. The reactive power limits of each PV inverter are strictly followed, with any excess set points automatically adjusted to the maximum allowed value, using Volt-VAR control.

The determination of the bus's operational status adheres to the rules outlined in the grid code DIN 50160 for medium voltage grid, also detailed in Table I and adopted from [22].

TABLE I. GRID CONSTRAINTS FOR MEDIUM VOLTAGE GRID.

| Grid constraints | Limits |
|---|---|
| Bus $\Delta$V must be within $\leq$ | 0.1 p.u./min |
| Loads to sustain $\Delta$V of $\leq$ | 0.02 p.u./min |
| Generators must endure $\Delta$V of $\leq$ | 0.05 p.u./min |
| Line load must be $\leq$ | 100 % |

### B. Reactive Power Management

*1) Building Experiments:* The challenge of finding reliable data is mitigated by using the Q-Controller (4) as the expert data source, validated by Ju et al. [23] for stability in reactive power control. The reactive power controller is implemented using the palaestrAI framework, with SAC chosen for its efficiency in continuous action spaces, outperforming DDPG and PPO [4]. BCO is selected for its simplicity. Each simulation spans one year, with data collected in 15-minute intervals, incorporating both a one-year training period and a subsequent one-year testing phase.

The primary objective is to construct an ANN architecture focused on reactive power control, utilizing data generated from the Midas simulation, used as the offline data, to enable BCO in a later stage. Before designing the ANN, the dataset is prepared using (4) to generate $q_{t+1}$ based on the $q_t$ and $V_t$ values. This equation, adapted from Ju et al. [23], provides stability and convergence in reactive power control and serves as the foundation for controlling reactive power set points in the grid.

$$q_{t+1} = [q_t - D(V_t - 1)]^+ \tag{4}$$

Using the simulated data, consisting of voltage ($V_t$) and reactive power ($q_t$) values, the effectiveness of (4) is evaluated by generating $q_{t+1}$ values and comparing them to the original $q_t$ values. This analysis spans all 14 buses over a year. The neural network is then developed using $V_t$[p.u.] and $q_t$[MVAr] as inputs ($x$) and $q_{t+1}$[MVAr] as the output ($y$).

14 datasets, ranging from 2500 to 34050 data points in increments of 2500, were generated from simulated data. For each dataset, a model is developed (e.g., $model_{2500}$ for 2500 data points, $model_{5000}$ for 5000 data points, etc.) with default hyperparameters initially, repeating this process across all datasets. The dataset is divided into training and testing sets, allowing for an assessment of model performance.

Along with generating 14 models for varying dataset sizes, hyperparameter optimization is performed using a randomized search method from Pedregosa et al. [24], to enhance performance of each model. This optimization targets five out of six hyperparameters—activation function, learning rate, number of neurons per layer, number of layers and batch size—while the number of episodes is fixed at 100. This choice is based on the observed trend in loss versus episode plots for the testing sets, where no significant reduction in loss occurs beyond 100 episodes. The evaluation of the models is based on mean, variance, Root Mean Square Error (RMSE) and Coefficient of Determination (R2). The model with the most favorable metrics—a mean similar to the original data, reduced variance, lower RMSE and higher R2—is selected for further experimentation.

These models are central to two of the three experiments explained later in Section III-B2, highlighting the importance of establishing a resilient network at this stage. Following this, three experimental approaches are explored: Supervised Experiment (SUP), SAC algorithm and a combination of SAC and ANN, thereby implementing BCO as depicted in Figure 2.

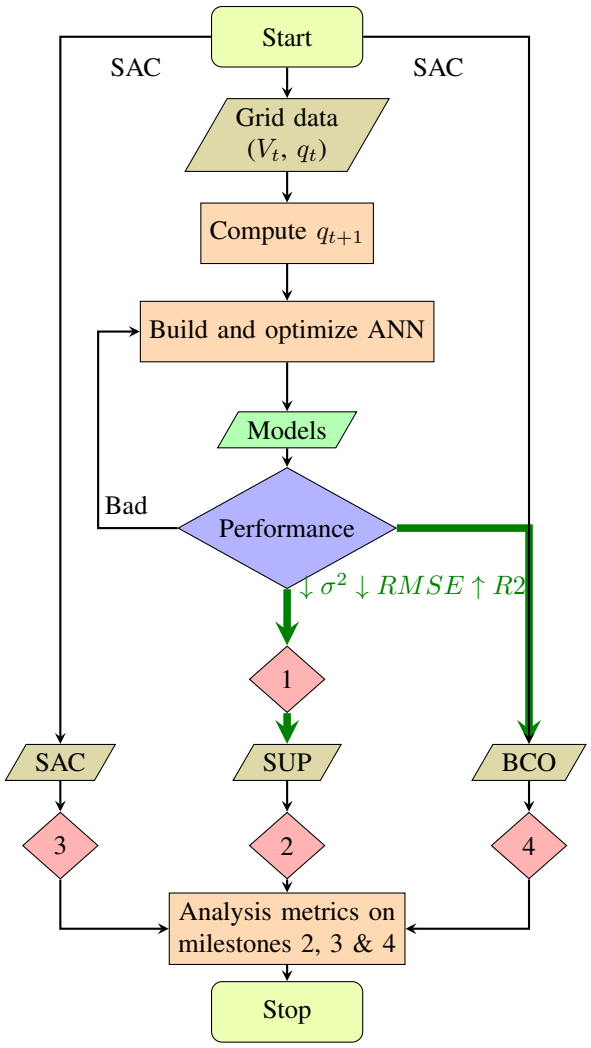



Figure 2. Process Flow Chart for Methodology.

*2) Experiment Setup:* The structure of the three experiments is presented in Table II. These experiments are designed to build progressively upon each other.

The first experiment, referred to as the SUP experiment, is the simplest and is based entirely on expert knowledge derived from (4), without any RL training process. This expert knowledge is transferred to the ANN using supervised learning. At each time step, sensor values $(V_t, q_t)$ are input into ANN, which produces an output used to set the value of $q_{t+1}$. Therefore, in this setup, the ANN acts as an actuator without policy initialization, interacting with a medium-voltage (MV) grid environment.

In the second experiment, the SAC algorithm, a well-established method in RL, serves as a baseline for comparison. SAC encourages exploration in RL and operates as an off-policy algorithm, leveraging prior knowledge. However, due to random policy initialization, the SAC algorithm can experience prolonged simulation times in the learning phase.

The third experiment addresses the limitations of both earlier approaches by combining the SAC algorithm with the ANN constructed from the expert knowledge, as the initial policy instead of a random initialization. The objective of this experiment is to integrate the characteristics of the two previous experiments.

This approach, known as BCO, mitigates the initial slow learning problem observed in the SAC experiment, as the ANN provides a more effective starting point. Consequently, the third experiment is expected to outperform both the SUP experiment, which lacks a RL simulation process and the SAC experiment, which suffers from random policy initialization, as illustrated in Table II.

The fourth experiment is an ablation study that compares the voltage profiles and reward values across all three previously mentioned experiments, considering both single bus and multiple bus scenarios.

TABLE II. DESCRIPTION OF AGENT CONFIGURATIONS FOR EACH EXPERIMENT ON THE MV GRID ENVIRONMENT.

| Experiment | SUP | SAC | BCO |
|---|---|---|---|
| Objective | Reward calculation | | |
| Sensor | $V_t$, $q_t$, %load & in-service status | | |
| Actuator | $q_{t+1}$ | | |
| Policy Initialization | None | Random | ANN |

*3) Performance Metrics:*

a) Models

In the SUP and BCO experiments, a total of 14 models, each trained on datasets of varying sizes, are analyzed to identify the model that achieves optimal performance with minimal data usage. Key evaluation metrics for the ANN models include mean, variance, R2 and RMSE. These metrics will guide the selection of sample-efficient models for both SUP and BCO experiments. In the SUP experiment, the models are used as actuators and in BCO, the models are used in policy initialization as mentioned in Section III-B2, experiment setup.

b) Experiments

The evaluation of experiments is based on six key metrics: voltage stability, adherence to voltage limits, high-reward performance, consistency, sample efficiency and robustness under controlling multi-bus scenarios.

(i) Both voltage and reward values are tracked over time to assess stability, compliance and overall performance.

(ii) Performance Consistency is ensured by repeating each experiment four times and comparing the outcomes across trials, focusing on voltage and reward stability.

(iii) Sample Efficiency is evaluated differently for SUP and SAC. In the SUP experiment, it is identifying the model that achieves the best performance with the least data. The model providing the reactive power set point is trained on various datasets of differing sizes (from 2500 until 35040 data points), with each training repeated four times to evaluate performance consistency. The selection of dataset sizes was intended to examine the impact of sample size on model performance and efficiency. In SAC experiments, sample efficiency is calculated using the following two criteria:

- The rate of change of reward:

$$\eta_s = \frac{dR}{dt} \tag{5}$$

- The area under the reward versus time plot.

(iv) Robustness is evaluated by simultaneously managing two buses, while other buses maintain a $\cos\phi = 0.9$ so the reactive power depends on the PV power injection, which significantly impacts grid dynamics. A model was developed using 35040 data points, incorporating four inputs—voltage and reactive power for each bus—and two outputs for the reactive power of both buses. This model enables the simulation of concurrent control of the buses. Scenarios are deemed robust when the voltage remains within the range $0.85 \leq V_{b,t}$[p.u.] $< 1.15$ for all buses $b$ and time $t$, since beyond these limits buses will get disconnected from the grid and when the rewards are within the range of $0.90 \leq R_{b,t}$[-] $< 1.00$. The percentage of values within these specified limits is calculated to assess robustness.

(v) Comparison of Experiments uses a rolling average analysis over ten days to compare voltage and reward outcomes, while sample efficiency is assessed through reward rate of change and Area Under the Curve (AUC) values as mentioned earlier.

## IV. RESULTS

This section presents the performance evaluation of different reinforcement learning approaches for voltage control in a power grid scenario. The effectiveness of the SUP, SAC and BCO methods is analyzed in maintaining voltage stability and optimizing rewards under both single and multi-bus control scenarios. The evaluation is conducted by measuring the

percentage of voltage and reward values that fall within the acceptable ranges of $0.85 \leq V_{b,t}$[p.u.] $\leq 1.15$ and $0.90 \leq R_{b,t}[-] \leq 1.00$, respectively, ensuring reliable grid operation. The results of these experiments are summarized in Tables III and IV, where the former illustrates the single bus performance while the latter demonstrates the robustness against controlling two buses.

### A. SUP Experiment

The sample efficient model for single bus scenario ensures $99.9\,\%$ of voltage and $97.0\,\%$ of rewards fall within the desired ranges as shown in Table III. Buses 5 and 11 are chosen for evaluation of robustness against controlling two buses simultaneously for all the three scenarios. $99.83\,\%$ of voltage values and $70.80\,\%$ of reward values are within the specified ranges.

### B. SAC Experiment

Since SAC uses random initialization and no models, model optimization is not applicable. Therefore, simple SAC training runs are carried out. For single bus scenario, SAC has $99.9\,\%$ of voltage values within $0.85 \leq V_{b,t}[p.u.] \leq 1.15$ and $47.0\,\%$ of reward values within $0.90 \leq R_{b,t}[-] \leq 1.00$, as shown in Table III. The robustness against controlling two buses is demonstrated by $100.00\,\%$ of voltage values and $70.59\,\%$ of reward values falling within their respective desired ranges, as seen in Table IV.

### C. BCO Experiment

The sample efficient model for single bus scenario ensures $99.9\,\%$ of voltage and $65.60\,\%$ of rewards fall within the desired ranges (see Table III). For the robustness of BCO experiment involving simultaneous control of two buses, $100\,\%$ of the voltage values and $81.56\,\%$ of the reward values fall within the respectively desirable ranges.

## V. Discussion

This section presents key findings from the experiments, focusing on the performance of the proposed approach across different evaluation metrics.

### A. SUP Experiment

The analysis based on voltage violations and reward performance concluded that $model_{5000}$ outperformed those trained on larger datasets, demonstrating the trade-off between dataset size and model efficiency. This shows the potential advantages of smaller datasets in achieving improved model performance and robustness against over-fitting.

During the robustness analysis, both buses exhibited significant performance issues, with voltage levels dropping to zero in 59 instances ($0.17\,\%$), violating the grid code requirements. Although the overall objective function remained high (ranging from 0.7 to 1) for most of the simulation period, these voltage drops adversely affected the reward, reducing it to approximately 0.5 in certain instances. In total, 10232 occurrences of rewards below 0.9 accounted for $29.20\,\%$ of the total time steps.

These sudden voltage spikes, observed in Figures 4a and 4b, primarily occurred during periods of high solar irradiation from April to September. These spikes are likely due to solar input exceeding the voltage limits set by the grid code. As shown in Figure 4c, the reward follows an opposing pattern to the voltage spikes, with the reward decreasing for high voltage deltas, in accordance with the grid code. Consequently, the current model's robustness is questioned, as it fails to maintain compliance with the voltage limits across multiple buses, despite achieving reward performance for $70.80\,\%$ of the total time steps. For the robustness analysis, a consistency check through multiple simulation runs was not conducted to minimize the overall number of simulations performed.

An overview of the evaluated criteria is provided in Table III. The values displayed are for the best performing models only.

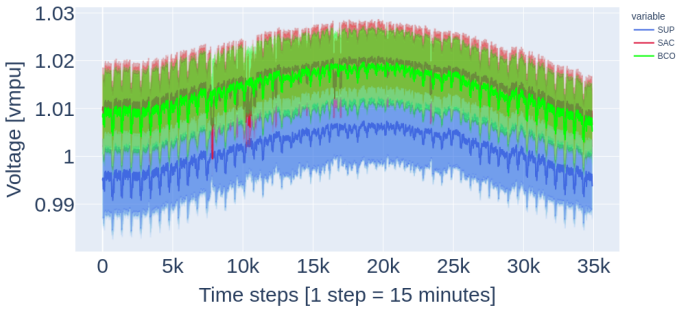
### B. SAC Experiment

In the single-bus scenario, the performance of the SAC experiment is the weakest among the three experiments evaluated. While the voltage values largely remain within the desired range, as presented in Table III, only $47.0\,\%$ of the reward values fall within the range $0.90 \leq R_{b,t}[-] \leq 1.00$, indicating poor adherence to the reward function. Considering the robustness of the SAC algorithm in multi-bus scenarios, it demonstrates impressive robustness. The average voltage performance of both buses, consistently maintains values between 0.96 and 1.03, with no grid code violations. Although reward values remain high, only $70.56\,\%$ of the values are between 0.9 and 1. This consistent voltage stability and moderately satisfactory reward performance demonstrate the resilience of the SAC algorithm (see Table IV).

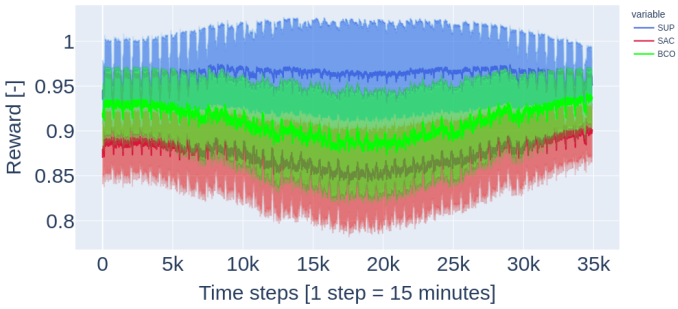
### C. BCO Experiment

Performance of $model_{30000}$ is the most sample-efficient performance based on voltage violations and reward values whereas the robustness analysis reveals that both buses exhibit higher voltages during sunny periods, likely due to increased solar irradiation, which affects reward acquisition but maintains grid code compliance. In contrast, voltages during the less sunny months remain within the desired range, with zero grid code violations observed. Overall, this scenario demonstrates a $0.17\,\%$ improvement in performance compared to the initial SUP experiment, suggesting that the ANN model, enhanced by the SAC algorithm, slightly outperforms the standalone SUP experiment.

The voltage distribution is primarily concentrated between 1 and 1.02 p.u., with some outliers between 1.02 and 1.06 p.u.. Interestingly, the reward performance reflects the inverse of the solar irradiation profile, with most values between 0.9 and 1 and $81.56\,\%$ falling within this range (see Table IV). This indicates a $10.97\,\%$ performance improvement for BCO over SAC, as both maintain $100\,\%$ compliance with voltage standards.




(a) Voltages of Single Bus.




(b) Rewards for Single Bus.

Figure 3. Comparison of Performances for Single Bus.

TABLE III. PERFORMANCE EVALUATION FOR SINGLE BUS

| Metrics | SUP | SAC | BCO |
|---|---|---|---|
| Voltage [%] | 99.9 | 99.9 | 99.9 |
| Reward [%] | 97.0 | 47.0 | 65.6 |
| Sample efficiency | | | |
| Data points | 5000 | N/A | 30000 |
| SAC algorithm | N/A | | |
| Slope | N/A | -0.0020 | -0.0024 |
| AUC | N/A | 22.45 | 22.75 |

TABLE IV. ROBUSTNESS EVALUATION FOR TWO BUSES

| Metrics | ANN | SAC | BCO |
|---|---|---|---|
| Voltage [%] | 99.83 | 100.00 | 100.00 |
| Reward [%] | 70.80 | 70.59 | 81.56 |

### D. Ablation Experiment

Table III represents a summary of the performance of single bus scenario for all the three experiments. The voltage occurrences within $0.85 \leq V_{b,t}[p.u.] \leq 1.15$ and reward occurrences within $0.90 \leq R_{b,t}[-] \leq 1.00$ are reported as percentages. Sample efficiency is measured by data points required for high model performance and by reward rate of change over the first 25 training hours for the SAC algorithm. Only the best-performing models' values are shown in this table.

*1) Rolling Average for Single and Multiple Bus Cases:* Figures 3a and 3b present the rolling voltage and reward averages respectively over the 34050 time steps for each experiment in the single bus case. The SUP experiment is constructed using 5000 data points model, while BCO experiment uses model generated from 30000 data points. In the multiple-bus case, Figures 4a and 4b show the voltage performances for Buses 5 and 11, respectively and Figure 4c demonstrates the reward performance. The rolling average approach smooths out outliers, enhancing the visibility of performance trends across time.

Figure 3a shows that all three experiments, SUP, SAC and BCO, maintain stable voltage performance. However, Figure 3b reveals that both SUP and BCO outperform SAC in reward collection, benefiting from the ANN model and initialization advantage, respectively. This model initialization, along with the high entropy inherited from SAC algorithm, consistently keeps BCO experiment ahead in reward collection compared to the SAC experiment. It is intriguing to observe that the performance of SUP experiment remains consistently higher than the other two experiments, making it the best performing experiment in the single bus case.

Figures 4a and 4b depict the voltage performances for Buses 5 and 11 in the multi-bus scenario. SUP shows frequent voltage drops, reflecting poor voltage management across both buses. However, both BCO and SAC experiments exhibit consistent voltage control, without any grid code violations. Short periods of voltage exceeding 1.02 p.u. reduce the rewards for SAC and BCO, but their overall performance remains excellent with performance improving with time. Figure 4c illustrates reward behavior for the multi-bus case, where BCO shows the best overall performance. SAC exhibits more resilient behavior than BCO during the high solar irradiation period, maintaining slightly higher rewards. BCO outperforms SAC in collecting more rewards during the lower solar irradiation period. Despite occasional voltage irregularities, BCO's performance highlights the effectiveness of using the ANN model for initialization, which consistently gives it a head start in reward collection, compared to SAC. Overall, BCO outperforms SAC by collecting $10.97\,\%$ higher rewards. SUP's performance, while strong in the single-bus case, falls short in multi-bus control scenarios.

*2) Sample Efficiency Comparison:* A comparison between SUP and BCO highlights the differences in sample efficiency in Table III. For the comparison of SAC and BCO sample efficiency, Figure 5 and Table V are used, as described in Section III, Methodology.

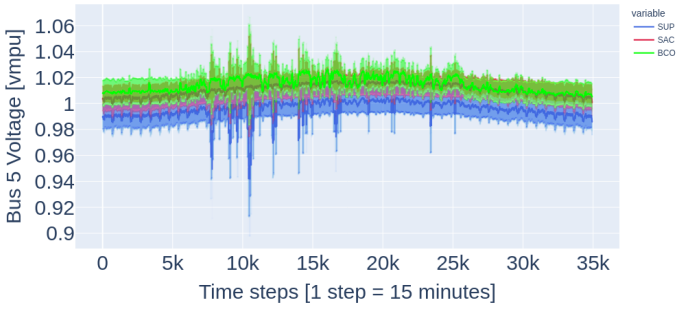
Referring to Table III, SUP, using only 5000 data points versus BCO's 30000, is more sample efficient, likely due to SAC's enhanced exploration capability. The complexity of BCO's combined model requires a larger dataset to capture patterns.

Considering (5), BCO shows a $20\,\%$ steeper slope and a $1.34\,\%$ larger AUC than SAC during the initial training phase (Figure 5), confirming its superior sample efficiency, summarized in Table V.

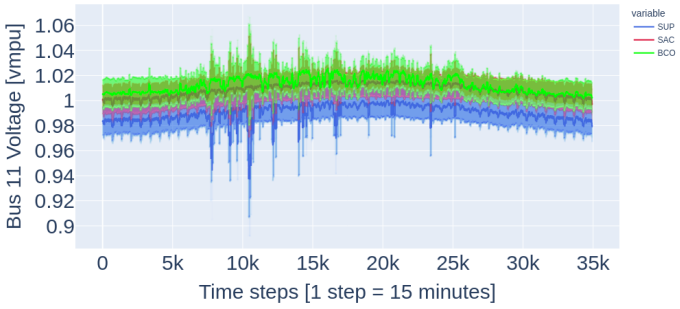
*3) Consistency and Robustness Assessment:* Consistency is evaluated by carrying out 4 repetitions for each experiment. Figures 3 and 4 show an average of these four repetitions.

TABLE V. COMPARISON OF SLOPE AND AUC IN THE FIRST 25 HOURS OF SAC AND BCO TRAINING.

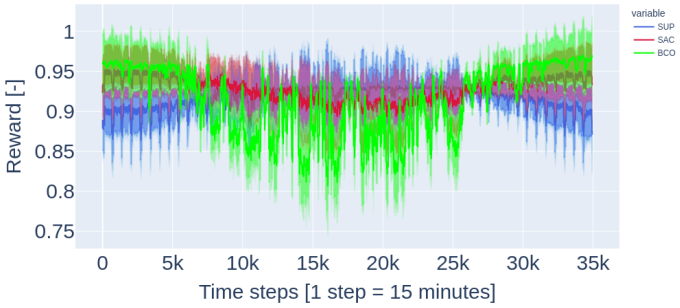
| Criteria | SAC | BCO | % Difference |
|---|---|---|---|
| Slope | -0.0020 | -0.0024 | 20.00 |
| AUC | 22.45 | 22.75 | 1.34 |




(a) Voltages of Bus 5.




(b) Voltages of Bus 11.




(c) Rewards for Two Buses.

Figure 4. Comparison of Performances for Two Buses.

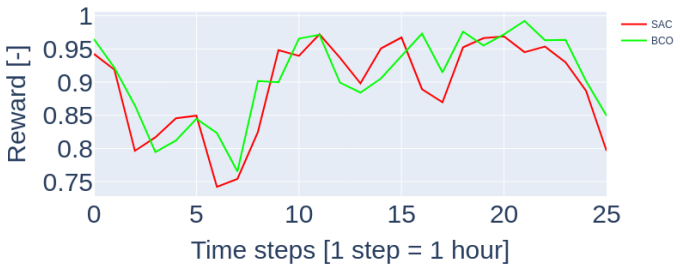



Figure 5. Sample Efficiency: Comparison of reward collection for initial 25 hours of training phase for SAC and BCO experiments.

Robustness against controlling two buses is evaluated based on the criteria mentioned in Section III, Methodology .

All three experiments demonstrate a certain level of consistency, although a minor inconsistency arises due to the distributional shift caused by employing different seed values for each of the four repetitions.

In terms of robustness in managing multiple buses, assessed by the percentage of occurrences where voltage and reward values stay within specified limits, BCO demonstrates the highest resilience. It is followed by SAC, which shows $10.97\,\%$ lower reward collection and then SUP, which, although gradually becoming more robust over time, exhibits lowest stability in multi-bus control due to frequent voltage violations and extended learning times.

## VI. Conclusions and Future Work

In the single-bus control scenario, the SUP experiment exhibits high sample efficiency for model by effectively utilizing a smaller dataset, achieving strong reward collection and stable voltage control. However, in multi-bus scenarios, it struggles with voltage stability and reward collection, bringing attention to limitations in handling more complex environments. Table VI can be referred for a summary of the results.

TABLE VI. OVERVIEW OF THE RESULTS.

| | Evaluation Criteria | SUP | SAC | BCO |
|---|---|---|---|---|
| Single Bus | Voltage Stability | ✓ | ✓ | ✓ |
| | Reward Collection | ✓ | × | - |
| | Sample Efficiency: Model | ✓ | | × |
| | SAC algorithm | | × | ✓ |
| Two Buses | Voltage Stability | × | ✓ | ✓ |
| | Reward gained | - | × | ✓ |

The SAC experiment demonstrates reliable voltage stability in both single- and multi-bus experiments but falls short in reward collection compared to the other methods, due to its weaker model initialization. Despite this, SAC's algorithm proves robust in managing complex grid conditions, although at the cost of sample efficiency.

The BCO experiment emerges as the best-performing method overall. It maintains superior voltage stability across both scenarios, shows excellent sample efficiency for the SAC algorithm and achieves higher reward collection, making it the most effective solution for both simple and complex grid control tasks.

Future research could enhance performance by exploring advanced neural network architectures and alternative hyperparameter optimization methods. Additionally, modifying the objective function to penalize voltage violations and conducting more repetitions could improve adherence to grid standards and strengthen result confidence.

## ACKNOWLEDGMENTS

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under grant AGenC (01IS22047C). The work on the ARL agent's architecture is funded by the BMBF under grant no. 01IS22071.

## REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[2] R. F. Prudencio, M. R. Maximo, and E. L. Colombini, "A survey on offline reinforcement learning: Taxonomy, review, and open problems", *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–8, 14, 2023.

[3] B. Kirby and E. Hirst, "Ancillary service details: Voltage control", Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), Tech. Rep., 1997.

[4] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies", in *International conference on machine learning*, PMLR, 2017, pp. 1352–1361.

[5] A. Nair, A. Gupta, M. Dalal, and S. Levine, "Awac: Accelerating online reinforcement learning with offline datasets", *arXiv preprint arXiv:2006.09359*, 2020.

[6] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor", in *International conference on machine learning*, PMLR, 2018, pp. 1861–1870.

[7] OpenAI, *Soft actor-critic (sac)*, https://spinningup.openai.com/en/latest/algorithms/sac.html, Accessed: 2025-02-13, 2023.

[8] Z. Ding and H. Dong, "Challenges of reinforcement learning", *Deep Reinforcement Learning: Fundamentals, Research and Applications*, pp. 249–272, 2020.

[9] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation", *arXiv preprint arXiv:1805.01954*, 2018.

[10] K. Chandrasekaran, J. Selvaraj, C. R. Amaladoss, and L. Veerapan, "Hybrid renewable energy based smart grid system for reactive power management and voltage profile enhancement using artificial neural network", *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 43, no. 19, pp. 2419–2442, 2021.

[11] C. Utama, C. Meske, J. Schneider, and C. Ulbrich, "Reactive power control in photovoltaic systems through (explainable) artificial intelligence", *Applied Energy*, vol. 328, p. 120 004, 2022.

[12] R. Fiorotti *et al.*, "A novel strategy for simultaneous active/reactive power design and management using artificial intelligence techniques", *Energy Conversion and Management*, vol. 294, p. 117 565, 2023.

[13] A. u. Rehman *et al.*, "Artificial intelligence-based control and coordination of multiple pv inverters for reactive power/voltage control of power distribution networks", *Energies*, vol. 15, no. 17, p. 6297, 2022.

[14] T. Wolgast and A. Nieße, "Learning the optimal power flow: Environment design matters", *Energy and AI*, vol. 17, p. 100 410, 2024, ISSN: 2666-5468. DOI: https://doi.org/10.1016/j.egyai.2024.100410.

[15] A. D'Silva, "Integrating behavioral cloning into a reinforcement learning pipeline", Available at https://www.politesi.polimi.it/handle/10589/208354, M.S. thesis, Politecnico di Milano, 2022.

[16] S. Dey, T. Marzullo, X. Zhang, and G. Henze, "Reinforcement learning building control approach harnessing imitation learning", *Energy and AI*, vol. 14, p. 100 255, 2023.

[17] S. Balduin, E. M. Veith, and S. Lehnhoff, "Midas: An open-source framework for simulation-based analysis of energy systems", in *International Conference on Simulation and Modeling Methodologies, Technologies and Applications*, Springer, 2022, pp. 177–194.

[18] A. Ofenloch *et al.*, "Mosaik 3.0: Combining time-stepped and discrete event simulation", in *2022 Open Source Modelling and Simulation of Energy Systems (OSMSES)*, IEEE, 2022, pp. 1–5.

[19] E. Veith *et al.*, "Palaestrai: A training ground for autonomous agents", in *Proceedings of the 37th annual European Simulation and Modelling Conference. EUROSIS*, 2023.

[20] E. M. Veith and E. Frost, "Cover me: Safeguarding multi-agent system with deep reinforcement learning for resilient grid operation", in *Proceedings of the 38th annual European Simulation and Modelling Conference*, Eurosis, Oct. 2024, pp. 3–4.

[21] T. Force *et al.*, "Benchmark systems for network integration of renewable and distributed energy resources", *no. April*, p. 63, 2014.

[22] E. Veith, T. Logemann, A. Wellßow, and S. Balduin, "Play with me: Towards explaining the benefits of autocurriculum training of learning agents", in *2024 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE)*, Dubrovnik, Croatia: IEEE, 2024, pp. 1–5. DOI: 10.1109/ISGTEUROPE56780.2023.10408277.

[23] P. Ju and X. Lin, "Adversarial attacks to distributed voltage control in power distribution networks with ders", in *Proceedings of the Ninth International Conference on Future Energy Systems*, 2018, pp. 291–302.

[24] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.