Leveraging Reinforcement Learning for Micro-grid Optimal Control

***Abstract*—The rise of renewable energy sources (RESs) is reshaping power grids, demanding new strategies for decentralized energy management. Microgrids (MGs) offer a solution by enabling localized energy control of generation, storage, and distribution. This paper proposes an approach based on reinforcement learning (RL) for optimizing MG energy management, where an RL agent learns optimal trading and storage policies using historical data. A digital twin (DT) simulates the dynamics of a battery energy storage system (BESS) for realistic emulation. Validated with real-world data from an Italian scenario, our method outperforms rule-based and existing RL benchmarks, providing a robust solution for intelligent MG management.**

I. Introduction

The energy landscape in Europe and the US has undergone a major transformation due to the widespread adoption of renewable energy sources (RESs). However, the decentralized nature of these energy sources challenges the traditional role of centralized power plants in maintaining grid stability. To effectively manage decentralized energy systems, new grid architectures such as micro-grids (MGs) have become essential. MGs are localized power networks capable of operating both in connection with the main grid and independently in *islanded* mode. Typically, they integrate multiple energy sources, including microturbines, fuel cells, and photovoltaic panels, alongside battery energy storage systems (BESSs) for efficient energy storage and distribution [1]. Even if MGs introduce a new level of flexibility, they also require tools for intelligent management, given the intrinsic combinatorial nature of the problem. A solution entails a distributed approach, where each MG independently defines its strategy for energy storage, purchasing from external sources, and selling surplus energy. In this direction, data-driven methods derived from the field of machine learning (ML), in particular, relying on deep reinforcement learning (DRL) techniques, have been proposed [2]. These methods have demonstrated the potential for learning optimal decision-making policies for MG control under conditions of uncertainty. Nonetheless, commonly available solutions optimize energy management based solely on costs and profits from MG interactions with the energy market while neglecting the BESS degradation and replacement costs [3], [4], [5]. A partial solution to these shortcomings is offered by [6], which models the system degradation only as influenced by the temperature, ignoring other important factors. The authors of [7] present a more sophisticated aging model. However, they assume constant energy market prices and static ambient temperature, neglecting that they significantly influence MG performance.

Instead, our work aims to design a more comprehensive controller that optimizes the MG to manage BESS usage, avoid excessive degradation, and reduce expenses due to energy exchanges with the main grid. In the current work, we propose:

- an RL-based [8] methodology to learn the optimal control strategy that uses historical information about energy consumption, power generation, market prices, and a digital twin (DT) to simulate the BESS;
- an experimental campaign to validate the proposed approach, which uses real-world data from the Italian energy market, real appliances consumption, photovoltaic production, and temperature profiles to build effective data-driven control policies through RL approaches.

II. Problem Formulation

An MG is a node connected to a broader grid E and comprises an overall generation source G, an overall consumption D, and a BESS B, as depicted in Figure 1. At a specific time t, power production coming from renewable generation sources (e.g., photovoltaic panels) generates an uncertain amount of power $P_{G,t}$. Meanwhile, the energy required to power the MG's electrical devices is aggregated within the total power demand $P_{D,t}$. The difference between production and consumption generates the net power:

$$P_{N,t} = P_{G,t} - P_{D,t}, \tag{1}$$

either positive or negative, which must be managed effectively by the controller. More specifically, if $P_{N,t}$ is a negative quantity, the controller has to decide whether to retrieve the remaining amount from the battery or the grid. Thus, it selects $P_{B,t}$, i.e., the fraction of $P_{N,t}$ that it wants to retrieve from the battery, while the remaining part $P_{E,t}$ is acquired from the main grid. Conversely, if $P_{N,t}$ is positive, it means that there is a surplus of power with respect to the users' demand. In this case, the controller action is intended as the proportion of net power $P_{N,t}$ to store within the battery $P_{B,t}$, while the remaining part $P_{E,t}$ will be sold to the grid. Formally, at each time t, the controller has to choose the value a_t with $a_t \in [0, 1]$ such that $P_{B,t} = a_t P_{N,t}$ and $P_{E,t} = (1-a_t)P_{N,t}$. The chosen action should comply with the BESS's physical constraints since we cannot overcharge or excessively drain it.

A. Controller's goal

The goal of the controller is to learn the optimal strategy leading to the maximization of economic profits due to trading with the energy market and the costs related to BESS degradation. Formally, the controller wants to maximize over a time horizon $\mathcal{T}$ the *total profit* obtained following a specific policy $\pi = (a_0, \ldots, a_{\mathcal{T}})$, defined as:

$$R_{\mathcal{T}}(\pi) = \max_{\pi} \sum_{t=1}^{\mathcal{T}} [r_{trad}(a_t) + r_{deg}(a_t)], \tag{2}$$

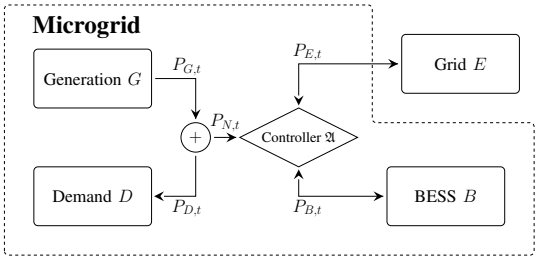

Fig. 1: MG schema.

where $r_{trad}(a_t) \in \mathbb{R}$ is the reward/cost gained from the exchanges of energy the MG makes with the market, and $r_{deg}(a_t) < 0$ is the cost due to battery degradation.

Formally, the trading component $r_{trad}(a_t)$ of the total profit is defined as:

$$r_{trad}(a_t) = (p_t^{sell} P_{E,t}^{+} + p_t^{buy} P_{E,t}^{-})\Delta\tau\,, \tag{3}$$

where p_t^{sell} and p_t^{buy} are the unit energy prices for selling and buying energy at time t, respectively (with $p_t^{sell} < p_t^{buy}$), $P_{E,t}^{+}$ and $P_{E,t}^{-}$ are the positive and negative part of $P_{E,t}$, respectively, and $\Delta\tau$ is the considered interval of time in $[h]$.

The degradation term $r_{deg}(a_t)$ depends on the BESS's State of Health (SoH), i.e., its aging. Indeed, an energy storage system starts its life cycle at the maximum SoH, i.e., 100%, which monotonically decreases until the end-of-life (EOL) value, indicating that the system is no longer capable of performing the current task (usually between $80 - 60\%$). Formally, having $\rho_t \in [0, 1]$, the degradation cost term is:

$$r_{deg}(a_t) = \frac{\rho_t - \rho_{t-1}}{1 - \rho_{EOL}}\mathcal{R}\,, \tag{4}$$

where ρ_t and ρ_{t-1}, with $\rho_t \leq \rho_{t-1}$, are the battery SoH values at time t and $t-1$, respectively, $\rho_{EOL} \in (0, 1)$ is the end of life (EOL) SoH, and $\mathcal{R}$ is the BESS replacement cost.

III. Proposed Method

This section formalizes the presented problem as a *Markov Decision Process* (MDP). In particular, we will define the state s_t, the action a_t, and the reward r_t we will use in our RL-based controller.

State space The state space vector s_t comprises the agent's signals from the environment. Formally, at each time step t:

$$s_t = \Big(\sigma_t, T_t, \widehat{P}_{D,t}, \widehat{P}_{G,t}, p_t^{buy}, p_t^{sell}, \cos(\varphi_t^d), \sin(\varphi_t^d), \cos(\varphi_t^y), \sin(\varphi_t^y)\Big), \tag{5}$$

where σ_t is the storage unit SoC, T_t is the battery temperature, $\widehat{P}_{D,t}$ and $\widehat{P}_{G,t}$ are the estimates of the energy demand and generation, p_t^{buy} and p_t^{sell} are the buying price and the selling price of the energy from the market, respectively, $\varphi_t^d \in [0, 2\pi]$ is a variable representing the angular position of the clock in a day, given by $\varphi^d = \frac{2\pi\tau_d}{\mathcal{T}_d}$, where $\tau_d \in [0, \mathcal{T}_d]$ is the current time of the day in seconds and $\mathcal{T}_d$ is the total number of seconds in a day and, similarly, $\varphi_t^y \in [0, 2\pi]$ is a variable representing the angular position of the current time over the entire year.

Action space The action space in our setting is the continuous action variable $a_t \in [0, 1]$, representing the proportion of energy to *dispatch* (*take*) to (from) the BESS. If $P_{N,t} > 0$, and therefore $P_{B,t} \geq 0$ and $P_{E,t} \geq 0$, it regulates the proportion between the power used to charge the battery and the one sold to the main grid. Conversely, if $P_{N,t} < 0$, the action a_t regulates the proportion of demanded energy that will be taken from the energy storage and the one bought from the market, thus $P_{B,t} \leq 0$ and $P_{E,t} \leq 0$.

Reward function Formally, the instantaneous reward is:

$$r_t = [r_{trad}(a_t) + r_{op}(a_t)] + \lambda r_{clip}(a_t). \tag{6}$$

The first two elements are the same as we defined in Equations (3) and (4). The term $r_{clip}(a_t)$ is a penalty given to the agent performing an action that needs to be clipped for structural constraints and λ is a weight to regulate how much such component should be influencing the learning procedure.

Algorithm 1 Interaction between Agent and Environment

Initialize: s_0, $\{\mathcal{P}_D^{(i)}\}_{i=1}^{M}$, $\mathcal{P}_G$, $\mathcal{C}_{buy}$, $\mathcal{C}_{sell}$, $\mathcal{K}$, $B(\cdot)$, $\pi(\cdot)$
for $j \in \{1, \ldots, n_{ep}\}$ **do**
  Sample demand profile $\mathcal{P}_D^{(i)} \sim Unif\big(\{\mathcal{P}_D^{(i)}\}_{i=1}^{M}\big)$
  Initialize σ_1, T_1, ρ_1
  for $t \in \{1, \ldots, \mathcal{T}\}$ **do**
    Compute estimates $\widehat{P}_{G,t}, \widehat{P}_{D,t}^{(i)}$
    Observe current state s_t
    Agent takes action $a_t \sim \pi(s_t)$
    Compute $P_{B,t} \leftarrow a_t(P_{G,t} - P_{D,t}^{(i)})$
    Update $(\sigma_{t+1}, T_{t+1}, \rho_{t+1}) \leftarrow B(\sigma_t, T_t, K_t, P_{B,t})$
    Compute $P_{E,t} \leftarrow (1 - a_t)(P_{G,t} - P_{D,t}^{(i)})$
    Collect reward r_t
    Update policy $\pi(\cdot)$

A. Learning procedure

The learning procedure is based on collecting samples (s_t, a_t, r_t, s_{t+1}) from the interaction between a learning agent $\mathfrak{U}$ and the environment. The pseudocode of the exchanges occurring during the training procedure is provided in Algorithm 1. The algorithm requires an initial state of the microgrid $s_0 \in S$, values for the time horizon $\{1, \ldots, \mathcal{T}\}$ of a set of demand profiles $\{\mathcal{P}_D^{(i)}\}_{i=1}^{M}$, with $M \in \mathbb{N}$, a generation profile $\mathcal{P}_G := (P_{G,1}, \ldots P_{G,\mathcal{T}})$, market selling and buying prices sequences $\mathcal{C}_{sell} := (p_1^{sell}, \ldots, p_{\mathcal{T}}^{sell})$ and $\mathcal{C}_{sell} := (p_1^{buy}, \ldots, p_{\mathcal{T}}^{buy})$ (respectively), and the ambient temperature profile $\mathcal{K} := (K_1, \ldots, K_{\mathcal{T}})$. We use multiple demand profiles where each specific profile is formally defined as $\mathcal{P}_D^{(i)} := (P_{D,1}^{(i)}, \ldots P_{D,\mathcal{T}}^{(i)})$. Conversely, we used a single generation, market price, and ambient temperature sequence, assuming the simulated MGs have similar locations. We also need a model for the BESS $B(\sigma_t, T_t, K_t, P_{B,t})$ which, given a SoC σ_t, the internal temperature T_t, the ambient temperature K_t, and a charge/discharge power $P_{B,t}$, is able to provide the SoC σ_{t+1}, its internal temperature T_{t+1} and the degradation occurred in the last step ρ_{t+1}. Finally, we require an initial policy $\pi(\cdot)$ that will be optimized over the procedure.

The learning process occurs over n_{ep} episodes, along which we simulate the MG by selecting each time a specific demand profile $\mathcal{P}_D^{(i)}$ uniformly at random, considered for the entire

training episode (Line 3). Once the episode is set and the battery is initialized (Line 4), we simulate a step t of the interaction between the environment and the agent. At first, the agent estimates the value of the next generation $\widehat{P}_{G,t}$ and demand $\widehat{P}_{D,t}$ (Line 6), to build the current state of the environment s_t (Line 7). Based on the state, the agent executes the action $a_t \sim \pi(s_t)$ prescribed by the policy (Line 8). Such an action determines the energy to be sent to the BESS $P_{B,t}$ (Line 9), and its corresponding evolution (Line 10), and the energy to exchange with the main grid $P_{E,t}$ (Line 11). Based on the degradation and the exchanges with the main grid, the agent receives the feedback r_t, as defined by Equation (6) (Line 12). Finally, the agent updates its policy $\pi(\cdot)$ (Line 13). Notice that while the environment is reset at each episode, the policy $\pi(\cdot)$ is kept fixed between episodes so that the information incorporated in previous episodes is retained.

IV. Experimental evaluation

We validate the presented approach in a real-world scenario based on an MG for household use in the Italian region. For the sake of brevity, we do not discuss the datasets used in this study; however, they will be made available in the associated GitHub repository. In our solution, denoted as $RL^\star$, the RL agent is trained using the PPO algorithm [9]. The training phase lasts $n_{ep} = 100$ episodes and has a decision step of $\Delta\tau = 3600s$ (or equivalently a frequency of 0.0003 Hz). We simulate the dynamics of the BESS using a DT, which emulates the evolution of battery SoC, temperature, and SoH over time. We assume a battery replacement cost of $\mathcal{R} = 3000$€.

We compared $RL^\star$ with common deterministic strategies and state-of-the-art RL-based controllers of MG energy storage systems. More specifically, we evaluated:

- *X-Y*s: deterministic rule-based strategies that dispatch $X\%$ of $P_{N,t}$ to the battery and the remaining $Y\%$ to the grid. Specifically, we tested 20-80, 50-50, and 80-20;
- *OnlyGrid (OG)*: a rule-based strategy forcing the interaction with the main grid without using the battery, corresponding to the 0-100 policy;
- *BatteryFirst (BF)*: a rule-based policy fostering battery usage as much as possible before interacting with the main grid, corresponding to 100-0 policy;
- *RL-base*: baseline reproducing the solution conceived by [7], trained with fixed ambient temperature and market prices, however adapting the method to a continuous action space.
- *RL-base+*: an extension of *RL-base* using fixed ambient temperature but incorporating information about the evolving market dynamics.

We compare the different policies $\mathfrak{U}$ in terms of average (over the validation profiles) empirical reward over the time horizon $t \in \{1, \dots, \mathcal{T}\}$:

$$\widehat{R}_t(\mathfrak{U}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{t}[r^{(i)}_{trad}(a_h) + r^{(i)}_{deg}(a_h)], \tag{7}$$

where $r^{(i)}_{trad}(a_h)$ and $r^{(i)}_{deg}(a_h)$ are the trading and degradation costs corresponding to the i-th demand profile $\mathcal{P}^{(i)}_D$ at time t. We also compared the separate contribution of trading and degradation to the final reward as follows:

$$\widehat{R}_{\square,t}(\mathfrak{U}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{t} r^{(i)}_{\square}(a_h) \tag{8}$$

with $\square \in \{trad, deg\}$. In particular, we evaluate the gap $\Delta\widehat{R}_{\square,t}(\mathfrak{U}, \mathcal{B})$ w.r.t. a baseline method $\mathcal{B}$:

$$\Delta\widehat{R}_{\square,t}(\mathfrak{U}, \mathcal{B}) := \widehat{R}_{\square,t}(\mathfrak{U}) - \widehat{R}_{\square,t}(\mathcal{B}). \tag{9}$$

Positive values for $\Delta\widehat{R}_{\square,t}(\mathfrak{U}, \mathcal{B})$ corresponds to larger performance of $\mathfrak{U}$ than $\mathcal{B}$.

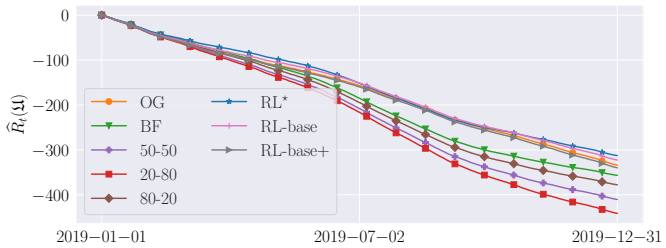

Fig. 2: Cumulative average empirical reward.

Campaign results Figure 2 reports the cumulative empirical reward $\widehat{R}_t(\mathfrak{U})$ of the policies corresponding to the analysed control strategies over the time horizon. It is worth noting that, on average, none of the strategies achieve a positive total return. This was expected since the cost of buying energy is always larger than the selling price, and the generation profile produces energy in excess of the demand only 39.6% of the times over the year of testing. Moreover, the reward also includes the degradation component, which is strictly negative since it is related to BESS usage. Among the analysed approaches, the proposed $RL^\star$ solution provides the best average results over the time horizon. Within the central months of the year (approximately June to September), the performances have been on par or slightly worse than those of previous RL-based solutions and OG. However, at the end of the time horizon, the $RL^\star$ method reduces the cost of 3.2% w.r.t. *RL-base*, i.e., the state-of-the-art RL solution, and of 6.7% w.r.t. OG, i.e., the best among the rule-based strategies. The strong OG performance, in this scenario, is primarily due to the relatively high cost of BESS replacement, which in turn brings high degradation costs for those methods that rely more heavily on BESS utilization.

Figure 3 presents a heatmap showing how $RL^\star$ plays different actions in response to varying demand levels throughout the test year. Each tile represents the logarithm of the number of times an action is chosen. This figure highlights that as the demand rises, the action progressively increases the amount of power the algorithm is addressing to the BESS. Using such a strategy, $RL^\star$ obtains a larger overall reward than OG. These results suggest that trivial rule-based approaches might not be flexible enough for this setting since they cannot adapt to external environmental changes.

Furthermore, we infer that *RL-base* and *RL-base+* tend to trade more with the market than $RL^\star$, and they make

less intensive usage of the BESS. This phenomenon is more evident for *RL-base+*, whose policy is almost equal to the OG. This might be because it employs approximate information about the degradation, overestimating the associated cost. This highlights how the joint use of information about the energy market and the ambient temperature, strongly influencing the degradation, is crucial to allow $RL^\star$ to learn effective strategies for the MG.

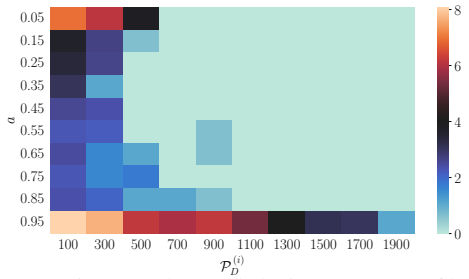

Fig. 3: Action vs. demand during a test profile run.

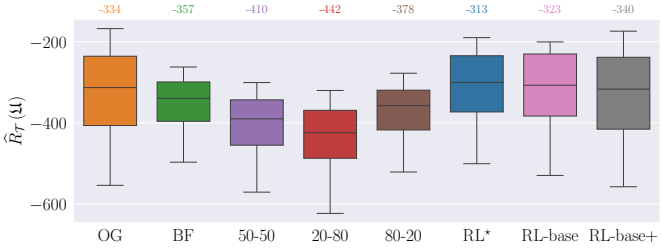

Fig. 4: Boxplot representing the empirical return across the 28 validation profiles. The values above the plot represent the mean over the test profiles.

In Figure 4, we provide boxplots representing the empirical distribution of the reward $\widehat{R}_{\square,t}(\mathfrak{U})$ at the end of time horizon $\mathcal{T}$ of the considered strategies over the different demand profiles. The figure shows that returns significantly fluctuate across testing profiles. This is because the yearly demand for different profiles spans from 1.50 to 5.05 MWh over the test set. However, we can see that the median of the reward for RL^* is close to the 75% percentiles of most rule-based algorithms and has lower variability w.r.t. the OG one. Even when compared with the other RL-based approaches, RL^* provides the largest median value for the reward and the smallest spread (having a smaller box and shorter whiskers). This corroborates the idea that the proposed approach can provide more consistent results over the considered profiles. Finally, a paired t-test to check if the difference between the rewards of RL^* and those of the other methods is significantly greater than zero provided us with a p-value (overall) of 0.0052, proving that there is strong statistical significance that our method is consistently performing better than the others.

Market shifts We evaluated the robustness of our approach across various scenarios with modified energy prices. In particular, we performed experiments in which we considered a multiplicative term $\alpha > 0$ for both the selling and buying prices, i.e., we used $\alpha\mathcal{C}_{sell}$ and $\alpha\mathcal{C}_{buy}$ in place of $\mathcal{C}_{sell}$ and $\mathcal{C}_{buy}$, respectively (during the training and testing of our algorithms). This modification mimics the increase ($\alpha > 1$) or decrease ($\alpha < 1$) of the energy price caused by geopolitical events or natural catastrophes.

Figure 5 show the results of the reward $\widehat{R}_{\mathcal{T}}(\mathfrak{U})$ with different values of the parameter $\alpha \in \{0.1, 0.5, 1, 1.5, 2\}$. When the energy cost is small, the most profitable strategy is the *OG* since the financial loss from market transactions diminishes. Conversely, as α increases, the optimal strategy shifts toward *BF* since purchasing energy becomes prohibitively expensive, making the battery degradation a more cost-effective choice. Moving to RL strategies, $RL^\star$ outperforms *RL-base* for $\alpha > 0.5$ and is on par with all the other approaches for all the values of α. These results corroborate that our approach can adapt to different environmental conditions.[1]

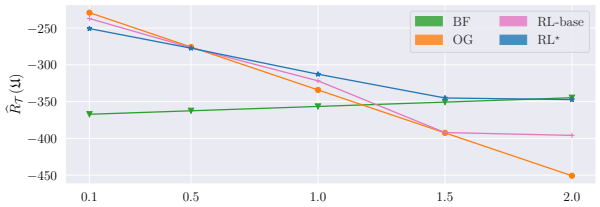

Fig. 5: Reward at the end time horizon for different parameter values α.

V. Conclusion and Future Works

In this paper, we explore MG energy management optimization, and, leveraging historical data and a BESS DT, we propose RL^*, a novel RL-based approach that learns cost-effective control policies. Validated on Italian MG settings, our method outperforms rule-based strategies and prior RL benchmarks. Future work includes integrating multi-agent logic for cooperative energy trading.

References

[1] L. Mariam, M. Basu, and M. F. Conlon, "Microgrid: Architecture, policy and future trends," *Renewable and Sustainable Energy Reviews*, vol. 64, pp. 477–489, 2016.

[2] N. F. P. Dinata, M. A. M. Ramli, M. I. Jambak, M. A. B. Sidik, and M. M. Alqahtani, "Designing an optimal microgrid control system using deep reinforcement learning: A systematic review," *Engineering Science and Technology, an International Journal*, vol. 51, p. 101651, 2024.

[3] D. Domínguez-Barbero, J. García-González, M. A. Sanz-Bobi, and E. F. Sánchez-Úbeda, "Optimising a Microgrid System by Deep Reinforcement Learning Techniques," *Energies*, vol. 13, no. 11, 2020.

[4] L. Liu, J. Zhu, J. Chen, and H. Ye, "Deep Reinforcement Learning for Stochastic Dynamic Microgrid Energy Management," in *2021 IEEE 4th International Electrical and Energy Conference (CIEEC)*, 2021, pp. 1–6.

[5] C. Guo, X. Wang, Y. Zheng, and F. Zhang, "Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning," *Energy*, vol. 238, p. 121873, 2022.

[6] Y. Sui and S. Song, "A multi-agent reinforcement learning framework for lithium-ion battery scheduling problems," *Energies*, vol. 13, no. 8, 2020.

[7] M. Mussi, L. Pellegrino, O. F. Pindaro, M. Restelli, and F. Trovò, "A Reinforcement Learning controller optimizing costs and battery State of Health in smart grids," *Journal of Energy Storage*, vol. 82, p. 110572, 2024.

[8] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018.

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

[1] Similar experiments have been conducted altering the battery replacement cost, and the obtained results align with the presented ones.