Comparison of different RL Algorithms for the energy management for bidirectional charging of EVs

Rainer Gasper

INES, Hochschule Offenburg
rainer.gasper@hs-offenburg.de

Abstract. This study compares different Reinforcement Learning (RL) Algorithms to optimize the bidirectional electric vehicle (EV) charging in a microgrid with dynamic pricing and renewable energy. The environment includes an EV, photovoltaic, stationary battery, household loads, and grid connection. The RL agents have to learn to maximize the profit of the microgrid under the constraint to reach the desired departure SOC's of the EV battery. The different RL algorithms are compared by their test results.

Keywords: Reinforcement Learning; DQN, LSTM, PPO, A2C; Micro Grid, Bidirectional Charging.

1 Introduction

The integration of electric vehicles (EVs) into microgrids presents a transformative opportunity for local energy systems. The bidirectional charging of EVs enables e.g. vehicle-to-grid (V2G) services, which leads to enhancing grid stability, reducing peak loads, and supporting renewable energy integration. However, managing bidirectional charging in microgrids is complex due to fluctuating energy demand, supply, and user behavior. Additionally, in case of bidirectional charging of the EV battery, the microgrid is equipped with an additional battery to the stationary battery and therefore this adds an additional degree of freedom in the actuation.

Reinforcement learning (RL) offers a promising strategy to address these challenges. By learning optimal policies through interaction with the environment, RL can manage charging and discharging decisions, adapting to real-time conditions and maximizing long-term rewards. In this article, we compare the Deep Q network (DQN) algorithm, with two different neural network architectures, as an off-policy strategy, with the Advantage Actor–Critic (A2C), and the Proximal Policy Optimization (PPO), as on-policy strategies, for maximizing the profit of a microgrid.

2 Related Work

The optimal energy management of microgrids with electric vehicles is a thriving research topic. This is underlined by the enormous number of research papers that have already been published and continue to be published. Therefore, for an overview of the literature we refer to the reviews [11], [12] and [13].

In [11], a systematic overview of RL applications for vehicle-to-grid (V2G) systems is given, which categorizes methodologies across power grid operators, aggregators, and EV users. It contains the evaluation of model-based, model-free, multi-agent, and Transformer-based RL techniques, emphasizing multi-agent frameworks for managing dynamic, large-scale V2G interactions. Crucially, the survey highlights how RL formulations address battery aging and

State of Charge (SOC) constraints by incorporating depth-of-discharge penalties and State of Health (SOH) tracking directly into optimization models. Finally, it identifies key open challenges, which are low sample efficiency, cyber-attack risks, and oversimplified degradation models, and points toward physics-informed and hierarchical RL as critical future research directions.

The review [12] categorizes RL techniques for EV dispatch, comparing single-agent and multi-agent frameworks across Grid-to-Vehicle (G2V), Vehicle-to-Home (V2H), and Vehicle-to-Grid (V2G) operation domains. It outlines critical open challenges, emphasizing the need for safe RL algorithms with operational constraint guarantees, scalable multi-agent architectures, and improved real-world deployment robustness against communication delays and noisy data.

In [13] the literature summary for model-free RL techniques for managing EV charging and discharging within smart power grids is given. It addresses grid stability challenges caused by unpredictable driver behavior, fluctuating electricity prices, and intermittent renewable energy integration. By evaluating various RL frameworks, architectures, and constraints, the review provides a comparative analysis of how RL approaches optimize energy management systems. Most of the cited approaches are DQN, Q-learning and SARSA, and so on, A2C or PPO are not mentioned.

A very similar work is presented in [1], it compares a wide range of RL algorithms for energy management. But no bidirectional charging of EVs is considered. Also, the action space differs, we restrict our agent to the battery powers in discrete steps.

3 Reinforcement Learning of Micro Grids

In RL, an agent interacts via actions a according to its policy π with an environment and learns the best actions given a scalar reward r and the state s of the environment. We are considering model free RL, and use neural networks as universal approximators. In the following we will describe the environment and the agents used in this work.

3.1 Environment

The environment consists of the microgrid and the reward function. The microgrid itself consists of an EV with bidirectional charging capability, photovoltaic as a renewable energy source, a stationary battery, and a household as a consumer load. Also, the grid connection point of the microgrid with dynamic electricity pricing is included, which acts as a source (buying) or sink (selling) for electric power. The reward is the sum of the negative total energy cost and a penalty to take into account the desired end SOC of the EV battery.

We formulate the energy management problem as a Markov Decision Process (MDP), following [1]. Also, we know that the microgrid is only a partially observable MDP, because of the unknown future energy prices and PV power production. Additionally, the arrival time of EV is unknown. For an investigation to the approximation of the MDP properties, we consider the DQN with LSTM [14].

Microgrid

The microgrid consists of a PV power source, a grid connection point, a stationary battery, and a household as a power consumer. Also, an EV with capability for bidirectional charging can be connected to the microgrid. This leads to the following microgrid power balance

$$P_{Grid} - P_{Load} + P_{PV} - P_{Batt}^{stat} - P_{Batt}^{EV} = 0$$

where P_{grid} is the power of the grid, P_{load} is power consumption of the household, and P_{PV} is the produced photovoltaic power. The different powers are explained in the following.

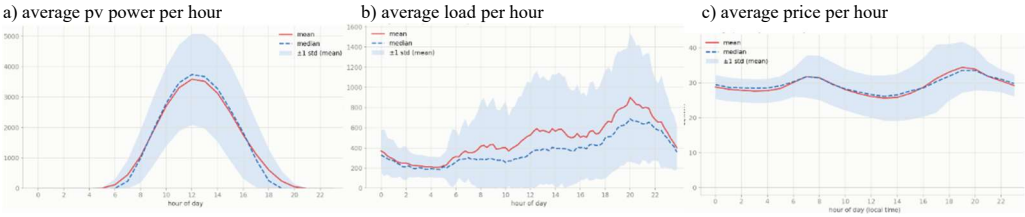


Figure 1: Statistical analysis of the PV power data, load, and dynamic price, mean and median values over the daily hour

PV Power

To train the agents on realistic data for the PV Power, we used the open source data from [6] for the year 2024 and for Germany. The data was scaled to match a 9kW peak photovoltaic system.

In Figure 1, a statistical analysis of the PV power is depicted. The mean and median PV power over daily hours show only a small difference. The curves show also the expected peak at 12 o clock. The maximal produced power is 6237 W. The total produced energy is 9519.2 kWh.

Household Power Consumption

For the load in the microgrid we assume a family household with 2 adults and 2 kids. The load profiles are generated with the synPRO tool [7] from Fraunhofer ISE.

In Figure 1, a statistical analysis of the load profile is given. Most of the loads are less than 1000 W. Over the year, only three times the load is over 4000W, with the maximal load of 4512 W. The total load energy is 4173.5 kWh.

Note, the accumulated load of the household is less than the generated PV power. At night there is no PV power production and the demand has to be met by the main grid or the battery powers. The main question is how the agents solve the optimization problem to maximize the profit of the microgrid and providing the desired departure SOC.

Grid Connection with Prices

To ensure the proper operation of the microgrid also if no PV or battery power is available, the microgrid is connected to the main grid. The main grid can instantly supply or accept power to or from the microgrid. The use of the main grid as power source comes at a price, which we assume consists of the spot market price plus a constant price for taxes, grid provider costs, and so on. We used the spot market prices in 2024 [8] for Germany and added additional costs for 2024 given in [9]. For the selling of power from the microgrid to the grid, we assumed a constant feed-in tariff of 8 ct/kWh, which holds for Germany.

A statistical analysis of the dynamic grid price is also shown in Figure 1. The minimal price was 3.85 ct/kWh and the maximal price 131.39 ct/kWh. The average price per hour of the day shows steady prices which are mostly governed by different taxes and so on in Germany.

Batteries

The storage dynamics of the stationary and the EV battery do not differ in the fundamental physical behavior, only in their parameters. Therefore, we give only one equation that describes

the battery. P_{batt} stands for the power of the stationary as for the EV battery power, respectively. The behavior of the battery is governed by the State of Charge (SOC), SOC ϵ [SOC_{min}, SOC_{max}],

$$SOC_{t+1} = SOC_t + \Delta t\, \eta(P_{Batt})\, \frac{P_{Batt}}{C}$$

where Δt is the time step, equal to 1h, and C the capacity of the battery. The efficiency η depends on the battery power P_{batt} and takes the losses into account. If P_{batt} is positive and we charge the battery, the efficiency is less than 1. Is P_{batt} negative and we discharge the battery, the efficiency is greater than 1. The efficiency depends also on the SOC and further quantities, but for simplicity we assume a constant efficiency for charging and discharging each.

The stationary battery is always available where the EV battery is only available if plugged into the microgrid. We assume typical a commuter behavior, so that the EV battery is not available in mean from 8 o clock to 17 o clock. The hours vary randomly in a normal distribution between 7 to 9 for departure and 16 to 18 for arrival. Also for trip energy of the EV we assumed that the EV SOC varies between 0.2 and 0.6 (normal distributed) at the arrival, in combination of a SOC = 0.8 at departure, the trip energy is high. No degradation of batteries is modelled, and is also not considered in the reward r.

Reward

We want to maximize the revenue of the microgrid, which is the negative costs of operating the microgrid. At the same time, we have to make sure, that the end-SOC of the EV equals the desired departure SOC for the EV. The two opposite goals have to be considered in the reward function

$$r = Rev - D^{EV}$$

where Rev is revenue and D^{EV} is the penalty if the SOC is not equal to desired SOC of the EV battery. The revenue is positive for feeding energy back into the main grid and negative for buying energy from the main grid

$$Rev = \begin{cases} -P_{Grid}\, \Delta t\, \mathrm{Pr}_{buy} & if \;\; P_{grid} \geq 0 \\ -P_{Grid}\, \Delta t\, \mathrm{Pr}_{sell} & if \;\; P_{grid} < 0 \end{cases}$$

where Pr_{buy} is the dynamic price and Pr_{sell} is the constant feed-in tariff, both described above. The penalty to enforce the desired SOC of the EV battery for departure is formulated as

$$\Delta SOC = \max\left(SOC_{des}^{EV} \frac{t - t_{av}}{t_{dep} - t_{av}} - SOC^{EV}, 0 \right)$$
$$D^{EV} = Pen\, \Delta SOC + (Pen\, \Delta SOC)^2$$

where SOC_{des}^{EV} is the desired departure SOC of the EV, t_{av} the start time when the EV is available for bidirectional charging, and t_{dep} is the departure time. Pen is a weighting factor. The penalty is only active if the SOC_{EV} is below the linear ramp up to the desired SOC_{des}^{EV}, is the SOC_{EV} higher than the linear function the penalty is equal to zero, which is ensured by the max function. The penalty forces the agents therefore gently to the desired SOC_{des}^{EV}, without constraining the usage of the battery if the SOC is higher than the linear boundary. Another alternative for the

penalty would be, that it is zero until a minimal time distance to the departure is reached, and the agent is forced to charge the battery with maximal battery power. In our experiments, the first depicted solution leads to more satisfying behavior of the agents, because the second case leads to sparse rewards.

Observation and action space

Now, we want to summarize the observation and action space. The observation space consist of actual values of the load power, the generated PV power, the SOC of the stationary battery, the EV battery SOC, the buying price for grid energy, and the actual time in daily hours. Also we added the departure time of the EV, but not the arrival time. The observation space is normalized so that all values are in the range of [0, 1] or [-1, 1] respectively.

The action space contains the EV battery power and the power of the stationary battery. The powers are discretized into 11 discrete steps each. In overall this leads to a combination space of 121, which can be still handled by DQN and so on. If the combination space increases further, the performances of the agents deteriorate because of the curse of dimension.

3.2 Agents

The agent interacts with the environment via actions depending on the actual states and the reward. The agent observes the states and chooses actions according to its policy, which is a mapping of the states to the actions. The goal of the agent is to maximize the expected return from each state. As strategies for the agent, we compare DQN [2] as an off-policy strategy, with A2C [5] and PPO [4], as on-policy strategies.

DQN combines Q-learning with deep neural networks to approximate the action-value function Q(s,a). The DQN is implemented with experience replay, where the transitions (state, action, reward, next state) are stored in a memory buffer and sampled randomly during training [2]. Also, a target network is used to compute target values [3]. Both techniques improve the learning process by reducing oscillations and breaking correlation between sequential data [2, 3]. For the exploitation and exploring trade-off, we implemented Ɛ-greedy-strategy with decreasing Ɛ.

DQN solves the Bellman equation by approximating the value function Q(s,a) with a deep neural network. For DQN, we compare different neural networks: a feedforward Multi-Layer Perceptron (MLP) , and LSTM, which leads to DRQN [14]. The internal memories of LSTM offer a more consistent learning and decision making, because the LSTM network depends not only on the actual state, which offers advantages if partial observable Markov Decision Processes (POMDP) are considered [14]. Note, that the described microgrid is POMDP, the prices, household load, and PV power generation depend on more than given in the current observations.

The A2C [5] approach uses two neural networks. The first neural network for the actor, which contains the policy, and observes the state of the environment and decides which action to take. The second network is the critic, it uses value functions to evaluate the actions taken by the actor. It calculates how much better (or worse) the outcome was compared to what it expected, the so called advantage.

In A2C [5] (and also in PPO) the advantage function A(s,a) is used to stabilize the learning. The advantage is positive if the Q-value is greater for action a than the expected value in state s, so an advantage is expected by choosing the action a, and vice versa. The advantage function is estimated by the function $\hat{A}_t = \sum_{i=0}^{k-1}(\gamma^i r_{t+i}) + \gamma^k V(s_{t+k};\theta_v) - V(s_t;\theta_v)$, where the value function V is approximated by a neural network with parameters θ_v.

PPO [4] is, like A2C, an on-policy actor-critic method. The policy update is restricted by a clipped surrogate objective, which prevents the agent from making too large steps in the policy update, which ensures smoother and more stable training. The objective function with the clipping function is given by

$$L^{clip} = \hat{\mathbb{E}}_t \left[\min \left(\frac{\pi(a_t|s_t;\theta)}{\pi(a_t|s_t;\theta_{old})} \hat{A}_t, clip\left(\frac{\pi(a_t|s_t;\theta)}{\pi(a_t|s_t;\theta_{old})}, 1-\epsilon, 1+\epsilon \right) \hat{A}_t \right) \right]$$

The clipping function constraints the probability ratio to the interval $[1-\epsilon, 1+\epsilon]$. Note, that the minimum depends on the sign of the estimated advantage. If $\hat{A}_t$ is positive, L^{clip} is constrained to $(1+\epsilon)\hat{A}_t$ if the probability ratio is higher than $1+\epsilon$, and if $\hat{A}_t$ is negative, L^{clip} is constrained to $(1-\epsilon)\hat{A}_t$ if the probability is smaller $1-\epsilon$. [4]

3.3 Training and configuration of the agents

To implement the agents for DQN, A2C and PPO we used the python package Stable-Baselines3 [10]. For the implementation of DQN with LSTM we had to implement our own version with pytorch [15]. The reason was that Stable-Baselines3 does not support sequence learning in DQN. By sequence learning [14] we are referring to the fact that we want that the agent learns the time dependence of the microgrid, where LSTM with the internal memories has its advantage.

We trained all agents on the data from year 2024 shown in chapter 3. The data was split into train data and test data by taking randomly 1 day out of each week for testing to account for the seasonality. All rewards reached a steady plateau during training.

For the DQN, A2C and PPO agents, we have chosen three layers with 256 neurons each. The activation function was ReLU. In case of the DQN with LSTM the network consists of 2 hidden layers with 256 LSTM units each. During the training of the LSTM-network the parameter gradients have been clipped to 1.0 to avoid great changes. Also we used the smooth L1 loss in pytorch [15] with beta = 1.0 for the objective function, which is also used in the DQN.

4 Test results

In the test, we compare the DQN, DRQN, A2C and PPO agents for maximizing the profit of a microgrid consisting of one household with an EV. For the comparison, we consider the baseline where the difference between produced PV power and load is sold or bought, depending on the signum, and the EV battery is charged as fast as possible when plugged in.

Table 1: Test results for all four agents, trained 5 times, compared to the baseline. Values are from 5 test runs for each agent.

Agent	Mean return	St. dev.	Min	Max	Mean € per Day	Mean SOC	St. dev	Below 0.8	Worst
A2C	-0.990	0.016	-1.019	-0.979	4.79	0.8023	0.0007	30.7%	0.755
PPO	-1.040	0.046	-1.102	-0.996	5.07	0.8028	0.0035	29.4%	0.738
DQN	-1.091	0.054	-1.159	-1.031	4.94	0.7941	0.054	42.4%	0.614
DRQN	-1.457	0.183	-1.726	-1.260	6.83	0.8002	0.0068	54.3%	0.713
Base-line	-2.105	-	-2.105	-2.105	10.45	0.8043	-	25%	0.783

The test results for 5 training runs is shown in Table 1. Looking at mean return, A2C is best. But comparing also standard deviation and the minimum and maximum returns, A2C and PPO are very similar in the numbers. DQN is near the maximal returns of A2C and PPO. DRQN is between DQN and the Baseline. The performance of DRQN was at first astonishing. For verification, we run DRQN with sequence length of 1 in the training and the results improved to the level of DQN. As in [14] stated, if the MDP requirements are nearly met, the recurrent behavior of LSTM has no advantage in the performance. But in our case, the LSTM increases the training effort. The results suggest, that the information in the observation space is sufficient so that the MDP formulation is adequate for this setup.

Besides maximizing the reward, the departure SOC has to be met by the agents. The resulting departure SOCs are analyzed and also depicted in Table 1. The mean value of the departure SOC are all over the desired value of 0.8, except for DQN, which is slightly below. Also DQN got the worst departure SOC of 0.614 in the test runs. DRQN missed with 54.3% the desired departure SOC the most.

Overall we can conclude that the higher training effort for DRQN does not pay off and DRQN is better than the baseline but not better than the more simple DQN without recurrent properties.

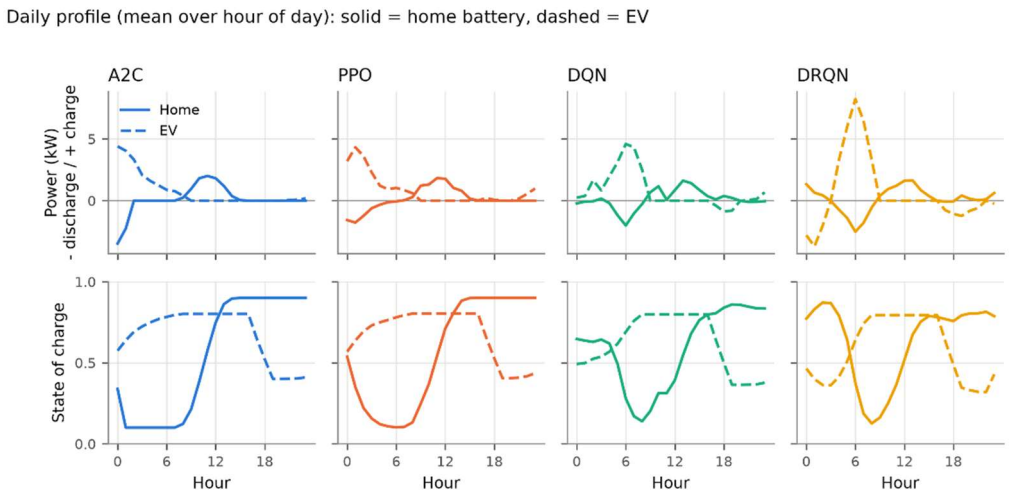


Figure 2: Daily power and SOC profile, mean from one test run of the 4 agents.

We consider the mean values of the battery powers and the SOCs over daily hours in Figure 2. Note, that the EV SOC is hold constant during the absence of the EV and the drop in SOC between 16 and 18 o'clock is the trip energy. The comparison shows that A2C and PPO are behaving very similar. They charge the home battery over the day when the maximum PV power is available, which also DQN and DRQN do. The EV is charged after midnight, where DQN and DRQN tend to charge at 6 o' clock in the morning. Taking the electric prices into account, Figure 1, A2C and PPO are charging at a lower price.

5 Conclusion and Outlook

The best results are obtained by A2C and PPO agents. The comparison of DQN and DRQN have shown, that no better performance is reached by adding recurrence to the agent.

The results could be improved by incorporation of predictions of the price, PV power and load. The degradation of the batteries is also very important, because the tradeoff between savings in the energy price has to be compared to the costs of additional degradation of the EV battery. In the development, the agents showed a degradation in performance if the discrete steps increase. This is due to the curse of dimensionality. Therefore for future work we plan to implement agents with continuous actions for power distribution and prosumer behavior.

Acknowledgement
R. Gasper acknowledges the financial support from the Carl-Zeiss-Stiftung.

References

1. Nakabi T. A., Toivanen P., Deep reinforcement learning for energy management in a microgrid with flexible demand, Sust. Energy, Grids and Networks, Volume 25, 2021, https://doi.org/10.1016/j.segan.2020.100413
2. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M., Playing Atari with Deep Reinforcement Learning. CoRR, 2013, https://doi.org/10.48550/arXiv.1312.5602.
3. Mnih, V., Kavukcuoglu, K., Silver, D. et al., Human-level control through deep reinforcement learning. Nature, 529–533 (2015). https://doi.org/10.1038/nature14236
4. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O., "Proximal Policy Optimization Algorithms," arXiv preprint arXiv:1707.06347, 2017.
5. Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., Kavukcuoglu, K.: Asynchronous Methods for Deep Reinforcement Learning, 2016, https://doi.org/10.48550/arXiv.1602.01783
6. Pfenninger, S., Staffell, I., Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data, Energy, Volume 114, 2016, Pages 1251-1265, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2016.08.060.
7. https://synpro-lastprofile.de/ , accessed June, 2026
8. https://www.smard.de/home , accessed June, 2026
9. https://strom-report.com/strompreis-zusammensetzung/ , accessed June, 2026
10. Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., Dormann, N., Stable-Baselines3: Reliable Reinforcement Learning Implementations, Journal of Machine Learning Research, 2021, 22, number 268, https://jmlr.org/papers/v22/20-1364.html
11. Xie, H., Song, G., Shi, Zh., Zhang, J., Lin, Zh., Yu, Q., Fu, H., Song, X., Zhang, H., Reinforcement Learning for Vehicle-to-Grid: A Review, Advances in Applied Energy, Volume 17, 2025, https://doi.org/10.1016/j.adapen.2025.100214.
12. Qiu, D., Wang, Y., Hua, W., Strbac, G., Reinforcement learning for electric vehicle applications in power systems: A critical review, Renewable and Sustainable Energy Reviews, Volume 173, 2023, https://doi.org/10.1016/j.rser.2022.113052.
13. Abdullah, H. ,Gastli, A., Ben-Brahim, L., Reinforcement Learning Based EV Charging Management Systems–A Review IEEE Access, 2021, doi: 10.1109/ACCESS.2021.3064354
14. Hausknecht, M., Stone, P., Deep Recurrent Q-Learning for Partially Observable MDPs, 2015, https://arxiv.org/abs/1507.06527
15. Paszke, A., Gross, S., Massa, Fr., Lerer, A., Bradbury, J., Chanan, G., Killeen, Tr., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., PyTorch: An Imperative Style, High-Performance Deep Learning Library, Advances in Neural Information Processing Systems 32, 8024—8035, 2019