Optimizing Battery Storage Systems in Energy Microgrids: A Reinforcement Learning Approach comparing multiple reward functions

Giorgia Ghione
Department of Electronics and Telecommunications
Politecnico di Torino
Turin, Italy
giorgia.ghione@polito.it

Vincenzo Randazzo
Department of Electronics and Telecommunications
Politecnico di Torino
Turin, Italy
vincenzo.randazzo@polito.it

Marco Badami
Department of Energy
Politecnico di Torino
Turin, Italy
marco.badami@polito.it

Eros Pasero
Department of Electronics and Telecommunications
Politecnico di Torino
Turin, Italy
eros.pasero@polito.it

***Abstract*—Battery Storage Systems (BSS) are increasingly utilized to enhance renewable energy consumption and operational stability in energy microgrids. However, the uncertainty characterizing renewable energy generation poses significant challenges in the optimal control of BSS. Multiple Reinforcement Learning (RL) approaches have been presented to solve this optimization problem. However, a comparison of different targets for the training of the RL systems is rarely performed. This work compares different reward functions that enable efficient BSS usage in the power plant of a transport hub while expressing the problem as a partially observable Markov Decision Process (POMDP). A Proximal Policy Optimization (PPO) algorithm is trained using reward functions derived from financial targets and BSS efficiency objectives. Results indicate that reward functions aligning BSS usage with market trends lead to superior performance compared to traditional earnings-based objectives. Furthermore, limitations regarding training episode numbers and reward normalization are identified, suggesting avenues for future research. This study contributes to advancing RL-based approaches for optimal BSS management in energy microgrid environments.**

***Index Terms*—Battery energy storage systems, Deep Reinforcement Learning, Proximal Policy Optimization, Energy Microgrids**

I. Introduction

A promising emerging solution to enhance renewable energy consumption, reduce energy costs, and improve operational stability for energy microgrids is Battery Storage Systems (BSS) [1]. In the industrial sector, the integration of renewable energy sources on a large scale, such as photovoltaic systems, has led to an increased application of energy storage technologies [2]. Nevertheless, the instability and uncertainty which characterize renewable energy generation and load demand cause significant difficulties in the operation of energy microgrids [3]. For this reason, the development of efficient energy management systems to properly schedule dispatchable resources such as BSS is necessary.

Multiple techniques have been proposed in the research literature to solve this energy management optimization problem. In particular, Reinforcement Learning (RL) has emerged as a technology that could significantly impact how BSS are controlled and managed [4]. In particular, multiple Deep Reinforcement Learning (DRL) approaches have been proposed. For example, [5] presented a RL-based method to control the state of charge (SOC) of a multi-electrical energy storage system utilizing the Deep Deterministic Policy Gradients (DDPG) algorithm. In [6] a neural network integrated reinforcement learning framework for the joint control of a manufacturing plant and on-site microgrid was proposed. [7] developed a multi-agent RL approach for the optimal control of an energy and manufacturing system including a stationary battery. In [8] a safe DRL method is presented to optimally control a renewable-based energy hub integrated with multiple energy while guaranteeing the satisfaction of physical constraints. [9] mapped a stochastic optimization model to a Markov Decision Process (MDP) considering uncertainties and proposed an improved DRL approach based on the soft Actor Critic (SAC) model.

The approaches presented in the literature exploit either financial targets or energy efficiency [4]. However, a comparison of different targets for the training of the RL systems is rarely performed. Additionally, these approaches model the control problem as an MDP, assuming the complete observability [10] of the system: however, this assumption rarely holds in real-

This publication is part of the project PNRR-NGEU which has received funding from the MUR – DM 352/2022. Dr. Randazzo acknowledges funding from the research contract no. 32-G-13427-2 (DM 1062/2021) funded within the Programma Operativo Nazionale (PON) Ricerca e Innovazione of the Italian Ministry of University and Research.

world environments, which are subject to uncertainties and errors.

This work presents a comparison of multiple reward functions for the training of a Proximal policy optimization (PPO) algorithm for the optimal scheduling of a BSS integrated in the power plant of a transport hub. The optimal control problem is expressed as a partially observable Markov Decision Process (POMDP) [11] in order to provide a realistic and robust approach to handle the partial observability of the system.

II. System description

The case study for this work is the smart grid of a transport hub located in the Lazio region in Italy. The power plant includes a photovoltaic (PV) station and a battery storage system (BSS) to store energy produced by the PV station for later use to satisfy the demand of the transport hub. The PV station has a nominal power equal to 22 MWh. The BSS has a nominal capacity equal to 10 MWh, a charging efficiency η_c equal to 0.90 and a discharging efficiency η_c equal to 0.90. One full charge or discharge of the BSS takes 1 hour, so it is possible to charge a maximum of 2250 kW and to discharge a maximum of 2250 kW every 15 minutes, taking the respective efficiencies into account. The battery is connected only to the PV station and the power system located in the facilities, thus it cannot exchange energy with the national grid. On the other hand, the smart grid can purchase energy from the national grid to fulfil the demand when the PV production and the BSS discharged power are not sufficient, and sell energy to the national grid when there is a surplus in the PV production which cannot be stored in the BSS.

III. Methodology

A. Dataset

The dataset which was collected includes the production data of the PV station, the load of the transport hub facilities and the wholesale reference price of electricity for the energy exchange with the national grid, called Single National Price (SNP). The dataset spans the period from the 1 January 2023 to the 31 December 2023. The day-ahead SNP data are publicly provided every day and were collected in the selected time period. The PV production data and the load data have a 15-minute granularity, while the SNP data have an hourly granularity: therefore, the SNP data of each hour were repeated four times to obtain a 15-minute granularity. The total number of data points in the dataset is 35036, since one hour of load data is missing, from 23:15 of 31 December 2023 to 00:00 of 1 January 2024. In the considered period, the 15-minute load of the transport hub ranges from 0 MWh to 6.91 MWh, while the 15-minute electricity price ranges from 0.001 €/kWh to 0.870 €/kWh. The dataset was split into a training and test dataset: the training dataset comprises the concatenated data of the first week of one month every other three months, while the test dataset comprises the remaining data: thus, the training dataset consists of 3360 data points, and the test dataset includes 31676 data points.

B. Objective function

The objective function is the earnings function defined in Equation 1.

$$Z = C + R \tag{1}$$

C is the cost of the power purchased from the grid and R is the revenue from selling the surplus energy to the grid. They are defined in the following equations:

$$C = E_{GP} * (SNP + spread_P) \tag{2}$$

$$R = E_{GS} * SNP \tag{3}$$

where E_{GP} is the energy which is purchased from the grid, E_{GS} is the surplus of energy which is sold to the grid, SNP is the SNP, $spread_P$ is the fixed purchase price equal to 0.935 €/kWh. The amount of energy exchanged with the grid E_{GEx} was calculated as follows:

$$E_{GEx} = E_{load} - E_{PV} - E_{BSS} \tag{4}$$

where h_t is the hour, E_{load} is the electric load, E_{PV} is the energy produced by the PV station, E_{BSS} is the energy exchanged with the BSS. Thus, if $E_{GEx} >= 0$ then $E_{GP} = E_{GEx}$ and $E_{GS} = 0$; if $E_{GEx} < 0$ then $E_{GP} = 0$ and $E_{GS} = |E_{GEx}|$. The operation and maintenance costs of the BSS were neglected.

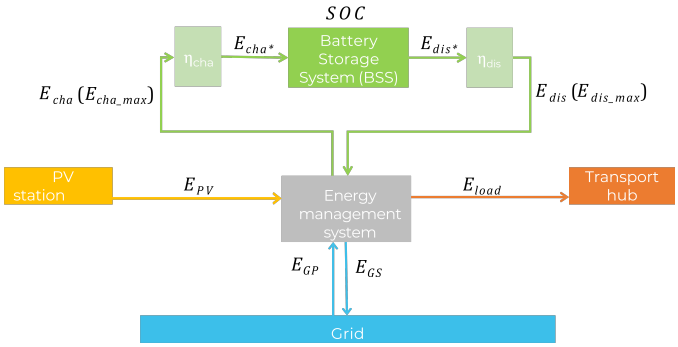


Fig. 1. Scheme of the energy management problem.

Figure 1 shows a summary of the energy management problem.

C. Problem formulation

Usually, RL problems are formulated as a Markov decision process (MDP), which is a discrete-time stochastic control process providing a framework to model decision-making problems. Three of the main components of the MDP are the state space S (the set of environment states that can be observed by the agent), the action space A (the set of actions that can be performed by the agent), the transition dynamics $T : S \times A \rightarrow \Delta(S)$, and the reward function $r = S \times A \times S \rightarrow \mathbb{R}$ (characterized by a discount factor γ). The

definition of MDPs explicitly assumes that the Markov property holds: it states that the distribution of the next state s_{t+1} depends only on the current state s_t and action a_t. However, the Markov property rarely holds in real-world environments, since the full state cannot be provided or includes uncertainty. Such problems are called partially observable problems [10] and they are commonly formulated as a partially observable Markov Decision Process (POMDP) [11]. A POMDP can be formally defined by a 6-tuple (S, A, P, R, Ω, O), where S, A, P, and R represent the MDP states, actions, transitions, and rewards, respectively. However, in a POMDP, the agent does not have access to the true system state and instead receives an observation $o \in \Omega$. This observation is derived from the actual system state based on the probability distribution $o \sim O(s)$.

In this context, the RL problem considered in this work is treated as a POMDP since the dynamics of the load and the PV production are not known and, thus, complete observability cannot be guaranteed. Specifically, the state s at each time t is defined as:

$$s_t \in S \tag{5}$$

$$s_t = (h_t, E_{load,t}, E_{PV,t}, E^{max}_{cha,t}, E^{max}_{dis,t}, SNP_t, SNP_{min}, SNP_{max}) \tag{6}$$

where h_t is the hour, $E_{load,t}$ is the electric load, $E_{PV,t}$ is the photovoltaic energy produced, $E^{max}_{cha,t}$ is the maximum amount of energy that can be charged in the BSS, $E^{max}_{dis,t}$ is the maximum amount of energy that can be discharged from the BSS, SNP_t is the predicted SNP, SNP_{min} is the minimum SNP of the current day, and SNP_{max} is the maximum SNP of the current day. The SNP for the current day is known one day in advance, thus these quantities can be calculated before starting the control procedure.

The observation o_t provided to the agent is a concatenation of the previous n states, i.e., $o_t = [s_{t-n}, s_{t-(n-1)}, \ldots, s_{t-1}]$, where n is a hyperparameter representing the length of the state sequence. This approach aims to capture sufficient historical information to approximate the hidden state of the environment, thereby enabling more informed decision-making. In this work, multiple values of n were tested (2, 4, and 6) and it was set to 2 since it is the value that yielded the best performance while minimizing the computational cost.

The action a at each time t is the amount of energy to be charged into the BSS or discharged from the BSS. It is defined as follows.

$$a_t \in A \tag{7}$$

$$a_t = E_{BSS,t} \tag{8}$$

When a_t is positive, the discharging energy is $E_{dis,t} = E_{BSS,t}$ and the charging energy is $E_{cha,t} = 0$; when it is negative, the charging energy is $E_{cha,t} = -E_{BSS,t}$ and the discharging energy is $E_{dis,t} = 0$.

Different reward functions r_t were tested. Their definition and description are listed below:

1) *Earnings with avoided cost*: the earnings computed in the objective functions in Eq. 1 considering also the saving from avoiding the cost of the power purchased from the grid when self-consuming the PV energy, denoted as AC, as defined in Eq. 10;
2) *BSS usage reward*: a function which positively rewards the correct usage of the BSS with respect to the SNP and penalizes inadvisable usage, as defined in Eq. 11;
3) *BSS usage reward and earnings 1*: the sum of the reward 2) and 3), multiplying 2) by a factor of $m = 100$, as defined in Eq. 12;
4) *BSS usage reward and earnings 2*: the sum of the reward 2) and 3), multiplying 2) by a factor of $m = 1000$, as defined in Eq. 12.

$$r_1(t) = C + R \tag{9}$$

$$r_2(t) = C + R + AC \tag{10}$$

$$r_3(t) = U \tag{11}$$

$$r_4(t) = C + R + AC + m * U \tag{12}$$

where SNP_{std} is the daily SNP standardized based on the minimum and maximum value for the current day, and the variables AC, $E_{SC,t}$, and U are defined in the following equations:

$$AC = E_{SC,t} * (SNP + spread_P) \tag{13}$$

$$E_{SC,t} = min(E_{load,t}, E_{pv,t} + E_{BSS}) \tag{14}$$

$$U = \begin{cases} \frac{E_{cha,t}}{E^{max}_{cha,t}} * SNP_{std} & \text{if } E_{BSS} < 0 \\ \frac{E_{dis,t}}{E^{max}_{dis,t}} * SNP_{std} & \text{if } E_{BSS} > 0 \\ \gamma = -\frac{SNP_{std}}{2} & \text{if } SNP_{std} > 0.5 \text{ and } E^{max}_{dis} \neg 0 \text{ or} \\ & \text{if } SNP_{std} < 0.5 \text{ and } E^{max}_{cha} \neg 0 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

D. DRL algorithm

The model selected in this work is PPO, which is a policy-based DRL algorithm based on an actor-critic architecture. This algorithm was first introduced in [12] as an improvement over standard policy gradient methods. The implementation utilized is the discrete version of PPO available in the OpenAI Spinning Up library [13], which was improved by adding Xavier weight initialization, gradient clipping, and a learning rate schedule.

The number of discrete actions chosen is 4751, ranging from -2250 kW to 2250 kW. The number of episodes set for training is 515.

E. Results and Discussion

To evaluate the models, a baseline was established: it was set to the earnings which could be obtained without the usage of the BSS, so self-consuming as much PV energy as needed to satisfy the demand, and exchanging energy with the national grid. This baseline was computed on the test set and it amounted to 5723786€.

TABLE I
SUMMARY OF RESULTS.

Experiment	Earning (€)
Baseline	5723786
Earnings with avoided cost	5762898
BSS usage reward	**5628387**
BSS usage reward and earnings 1	5733438
BSS usage reward and earnings 2	5744919

As described in III-C, the PPO algorithm was trained five separate times using different reward functions. In each experiment, the same training and test datasets were used, as well as the same hyperparameters and initialization of the PPO algorithm.

Table I summarizes the earnings of each model at test time. It is possible to observe that the model which performs best is the *BSS usage reward* one, providing a higher earning than the baseline. All the other models performed worse than the baseline. A reason for these results may be the fact that the reward function based on the trends of the SNP encourages the agent to exploit the BSS and hints at the most favourable moments of the day to charge and discharge the batteries. Thus, this reward enables a quick convergence of the model. On the other hand, since the reward function *Earnings with avoided cost* includes the earnings and costs from the exchange with the grid and the self-consumption savings, may be directing the model towards a maximisation of the energy sold to the grid, thus causing a minimal usage of the BSS. Such behaviour is evident when observing Fig. 2, which shows the *BSS usage reward* scenario over four example days (from 8 to 11 January 2023) in the sub-figure (a) and the *Earnings with avoided cost* scenario in the sub-figure (b) over the same days.

However, another factor which influences the results is the number of episodes. In particular, in this work, the number of episodes was set to a small number, i.e. 515, as a trade-off with regard to the consumption of computational resources. Nevertheless, the models with an earning-based reward may need a higher number of episodes to converge to an optimal solution, making the presented one a sub-optimal one. Another factor which may be coming into play is the reward magnitude: since the *BSS usage reward* is smaller on average than the other ones, it may enable more stable gradients and, thus, better performance. For this reason, reward normalization should be implemented. Such experiments are under development for the extension of this paper.

IV. Conclusion and Future Work

This work presents a comparison of multiple reward functions for the training of a discrete PPO algorithm for the optimization of the usage of a BSS in the facilities of a transport hub. The optimization problem was formulated as a POMDP in order to take into account the partial observability of the system. The reward function which yielded the highest earnings is a function which rewards a usage of the BSS which correctly follows the trends of the SNP: it enables the quick convergence of the model and it is superior to the baseline result. A reason for these results is the direct link of the reward function to the behaviour of the BSS, making the desirable actions more easily learnable to the agent. However, a limitation of this work is the small number of episodes for training and the lack of experiments regarding reward normalization. Another gap is the comparison of the PPO with innovative hybrid techniques for POMDP modelling, such as the combination of DRL with sequence-to-sequence encoder algorithms to extract a more meaningful representation of the state of the environment. Future work will additionally include the comparison of PPO with other discrete and continuous DRL models, such as Soft Actor-Critic and Deep Deterministic Policy Gradient.

Acknowledgment

A special thanks to Pellerey Valeria, Porro Gioele and Chiurato Simona of Trigenia S.r.l. for their support of this work.

References

[1] C. Zhao, P. B. Andersen, C. Træholt, and S. Hashemi, "Grid-connected battery energy storage system: a review on application and integration," *Renewable and Sustainable Energy Reviews*, vol. 182, p. 113400, 2023.

[2] S. Nyamathulla and C. Dhanamjayulu, "A review of battery energy storage systems and advanced battery management system for different applications: Challenges and recommendations," *Journal of Energy Storage*, vol. 86, p. 111179, 2024.

[3] M. M. Rana, M. Uddin, M. R. Sarkar, G. Shafiullah, H. Mo, and M. Atef, "A review on hybrid photovoltaic – battery energy storage system: Current status, challenges, and future directions," *Journal of Energy Storage*, vol. 51, p. 104597, 2022.

[4] R. Subramanya, S. A. Sierla, and V. Vyatkin, "Exploiting battery storages with reinforcement learning: A review for energy professionals," *IEEE Access*, vol. 10, pp. 54484–54506, 2022.

[5] F. Sanchez Gorostiza and F. M. Gonzalez-Longatt, "Deep reinforcement learning-based controller for soc management of multi-electrical energy storage system," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5039–5050, 2020.

[6] W. Hu, Z. Sun, Y. Zhang, and Y. Li, "Joint manufacturing and onsite microgrid system control using markov decision process and neural network integrated reinforcement learning," *Procedia Manufacturing*, vol. 39, pp. 1242–1249, 2019. 25th International Conference on Production Research Manufacturing Innovation: Cyber Physical Manufacturing August 9-14, 2019 — Chicago, Illinois (USA).

[7] M. Roesch, C. Linder, R. Zimmermann, A. Rudolf, A. Hohmann, and G. Reinhart, "Smart grid for industry using multi-agent reinforcement learning," *Applied Sciences*, vol. 10, no. 19, 2020.

[8] D. Qiu, Z. Dong, X. Zhang, Y. Wang, and G. Strbac, "Safe reinforcement learning for real-time automatic control in a smart energy-hub," *Applied Energy*, vol. 309, p. 118403, 2022.

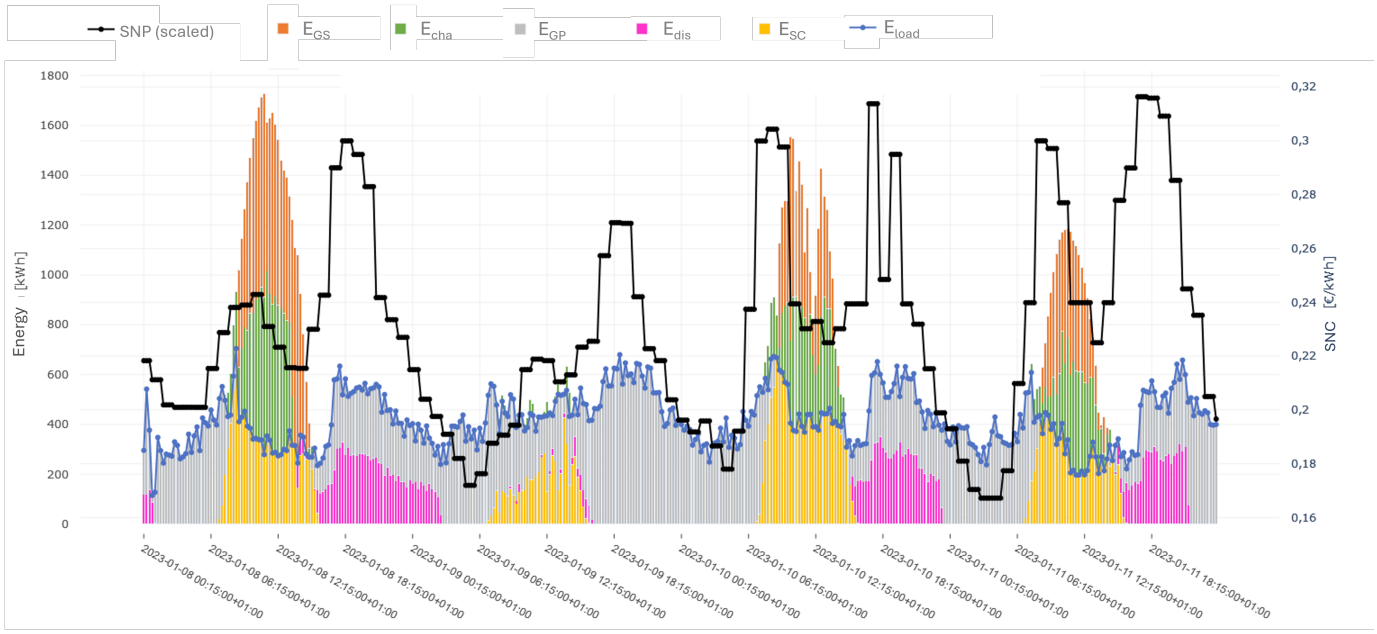
[9] Y. Zhou, Z. Ma, J. Zhang, and S. Zou, "Data-driven stochastic energy management of multi energy system using deep reinforcement learning," *Energy*, vol. 261, p. 125187, 2022.

[10] K. Åström, "Optimal control of markov processes with incomplete state information," *Journal of Mathematical Analysis and Applications*, vol. 10, no. 1, pp. 174–205, 1965.

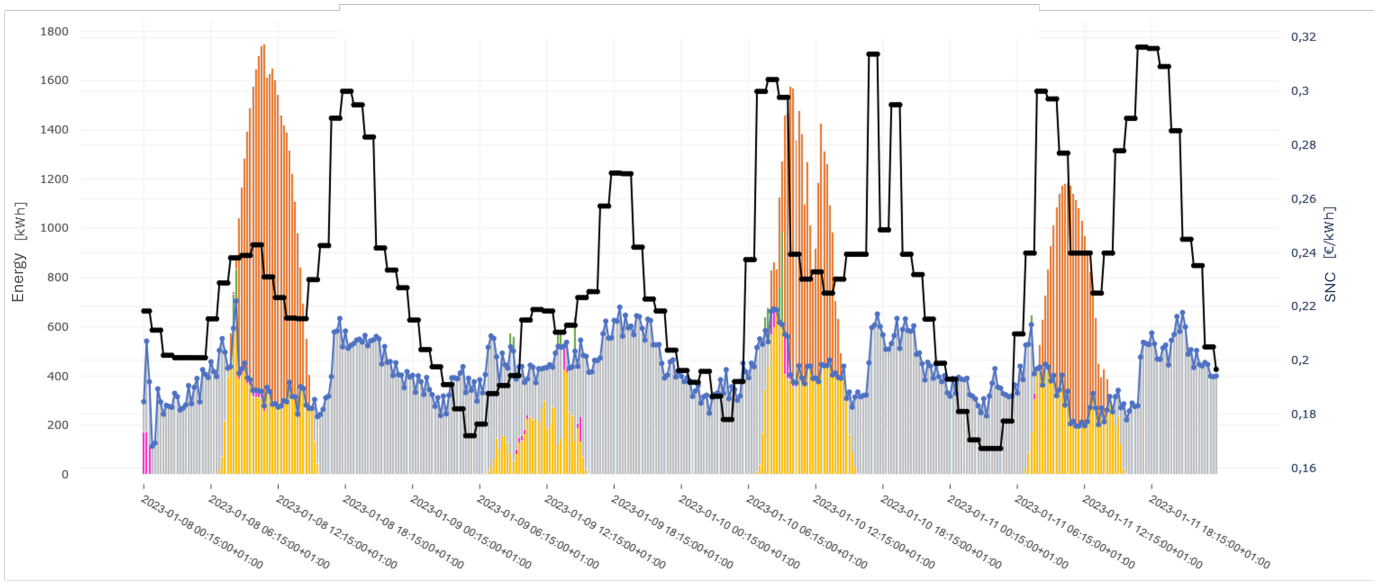
[11] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, no. 1, pp. 99–134, 1998.

[12] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017.

[13] OpenAI, "Welcome to spinning up in deep rl! — spinning up documentation." https://spinningup.openai.com/en/latest/, Accessed 2024. Online.



(a)



(b)

Fig. 2. Optimal BSS management in the *BSS usage reward* scenario (a) and in the *Earnings with avoided cost* scenario (b) over four example days, i.e. from 8 to 11 January 2023.