energies

Article

Optimal Energy Management of a Grid-Tied Solar PV-Battery Microgrid: A Reinforcement Learning Approach

Grace Muriithi and Sunetra Chowdhury *

Department of Electrical Engineering, University of Cape Town, Cape Town 7701, South Africa; MRTGRA010@myuct.ac.za

* Correspondence: Sunetra.Chowdhury@uct.ac.za

Abstract: In the near future, microgrids will become more prevalent as they play a critical role in integrating distributed renewable energy resources into the main grid. Nevertheless, renewable energy sources, such as solar and wind energy can be extremely volatile as they are weather dependent. These resources coupled with demand can lead to random variations on both the generation and load sides, thus complicating optimal energy management. In this article, a reinforcement learning approach has been proposed to deal with this non-stationary scenario, in which the energy management system (EMS) is modelled as a Markov decision process (MDP). A novel modification of the control problem has been presented that improves the use of energy stored in the battery such that the dynamic demand is not subjected to future high grid tariffs. A comprehensive reward function has also been developed which decreases infeasible action explorations thus improving the performance of the data-driven technique. A Q-learning algorithm is then proposed to minimize the operational cost of the microgrid under unknown future information. To assess the performance of the proposed EMS, a comparison study between a trading EMS model and a non-trading case is performed using a typical commercial load curve and PV profile over a 24-h horizon. Numerical simulation results indicate that the agent learns to select an optimized energy schedule that minimizes energy cost (cost of power purchased from the utility and battery wear cost) in all the studied cases. However, comparing the non-trading EMS to the trading EMS model operational costs, the latter one was found to decrease costs by 4.033% in summer season and 2.199% in winter season.

Keywords: microgrid; renewable energy resources; Markov decision process; Q-learning; smart control policy

Citation: Muriithi, G.; Chowdhury, S. Optimal Energy Management of a Grid-Tied Solar PV-Battery Microgrid: A Reinforcement Learning Approach. *Energies* **2021**, *14*, 2700. https://doi.org/10.3390/en14092700

Academic Editor: Gianfranco Chicco

Received: 19 March 2021
Accepted: 5 May 2021
Published: 8 May 2021

1. Introduction

Increasing interest in renewable energy sources has led to massive deployment of microgrids as they offer a scalable way of integrating renewable sources into the main grid while allowing maximum usage of battery energy storage system. In the long run, installation of microgrids are expected to reduce cost of power, dependency on utility grid, and increase rural electrification [1]. Nonetheless, increased integration of distributed renewable energy raises significant challenges in the stable and economic functioning of the microgrid as they are extremely volatile and random. These multiple stochastic resources combined with the load demand make preparation of accurate generation schedules very challenging. Deploying a battery energy storage system (BESS) [2] can significantly buffer the impacts of these uncertainties as it provides various auxiliary services to the power system i.e., load shifting, frequency regulation, voltage support and grid stabilization [3]. However, for a microgrid to guarantee reliable supply of power and efficient utilization of the battery storage, an energy management system (EMS) needs to be developed to optimally dispatch and distribute these energy resources based on their availability and associated costs.

Optimal energy management (OEM) involves the management/scheduling of various power system variables, in a day ahead context, in order to satisfy the load demand at minimal or acceptable costs while satisfying all technical and operational constraints. The main goal of developing an effective EMS is to achieve different objectives such as levelling peak loads, balancing energy fluctuations, maximizing renewable energy usage, reducing power losses, and increasing system load factor among others [4]. The EMS faces significant challenges as a result of the microgrid's existence, including small size, DRES volatility and intermittency, demand uncertainty, and fluctuating electricity prices. More advancements in microgrid design and control are needed to address these obstacles. To balance the high volatility of DRESs, additional sources of flexibility must be utilized at the architectural level. Furthermore, new, and intelligent control mechanisms are required to optimize energy dispatch and overcome microgrid's uncertainties.

Aimed at maximizing energy usage or reducing operational cost by managing intelligently the different types of energy resources and controllable loads in a grid-tied microgrid, several control approaches have been proposed. For years, conventional techniques such mixed-integer linear programming, linear programming, and dynamic programming have been proposed to optimally manage energy in microgrids [5–7]. These methods, however, are reported to suffer from the famous curse of dimensionality and are highly susceptible to getting sub-optimal results in environments that are highly stochastic, i.e., they contain volatile variables such as load demand, grid tariffs and renewable energy. Such techniques, therefore, have limited flexibility and scalability. Further, metaheuristics techniques including particle swarm optimization (PSO), genetic algorithm (GA), and their hybrids have also been used in literature to tackle the issue of energy management in microgrids [4,8–10]. However, these techniques involve extensive computational time and hence, they cannot be executed online. Online operation allows computing resources to be used more economically as it doesn't require one to have another committed computer for performing the optimization process offline. The aforementioned algorithms also don't have a learning component, i.e., they are incapable of storing the optimization knowledge and reusing it for a new optimization task [11]. Given that the load demand varies on hourly basis, it is required to calculate the schedule for every new generation and demand profile, and this is not computationally efficient. In addition, the performance of this techniques may deteriorate if accurate models or appropriate state variables forecasting are unavailable. Often, metaheuristic methods are hybridized with other linear methods for an advantage complementation. A comprehensive review of these decision making strategies and their methods of solution has been presented in [12,13].

In the last decade, intelligent learning-based techniques have made major progress in decision-making problems and have also proved ideal in overcoming these limitations, as they can automatically extract, monitor, and optimize generation and demand patterns. Additionally, they are capable of relaxing the idea of an explicit system model to ensure optimal control. This is of great benefit, since the problem of energy management is normally a partly observable problem, i.e., hidden or unknown information always exists.

The reinforcement learning (RL) method, one of the machine learning algorithms, is well known because of its ability to solve problems in stochastic environments. It aims at making optimal time-sequential decisions in an uncertain environment. Reinforcement learning involves a decision maker (agent) that learns how to act (action) in a particular situation (state) through continuous interaction with the environment so as to maximize cumulative rewards [14,15]. In the learning process, the agent is in a position to learn about the system and to take action that affects the environment so as to achieve its objective. In RL, the agent considers the long-term reward, instead of simply getting the immediate maximum reward. This is very important for resource optimization problems in renewable powered microgrids, where supply and demand are changing rapidly. Q-learning, one of the RL methods, is commonly used to solve sequential decision-making problems as explained by authors in [16]. Q-learning is an off-policy algorithm that doesn't require any prior knowledge of rewards or state transition probabilities of a system, thus making it applicable to systems that manage real-time data. Many scholars, focusing on microgrid EMS [11,17–19], specifically have used Q-learning to control energy. The key benefit of RL techniques is their adaptability to stochastic systems and ability to transfer knowledge, i.e., the information gained when learning policies for a specific load demand can be retrieved to learn an optimal schedule for other load profiles [11].

Taking advantage of these characteristics, several scholars have used this approach to solve the microgrid energy management problem. For instance, Brida et al. [20] used batch reinforcement learning to implement a microgrid EMS that optimizes battery schedules. Charge and discharge efficiency of the battery and the microgrid nonlinearity caused by inverter efficiency were considered. Elham et al. [21] presented a multi-agent RL method for adaptive control of energy management in a microgrid. The results indicate that the grid-tied microgrid learned to reduce its dependency on the utility grid significantly. Authors in [22] presented an optimal battery scheduling scheme for a microgrid energy management. A Q-learning technique is implemented to reduce the overall power consumption from the utility in [22] and simulation results show that algorithm reduces dependency on the main grid. However, this work fails to consider battery trading with the utility and the impact of battery life cycle from those actions. In [23] Zeng et al. suggested an Approximate Dynamic Programming (ADP) method to tackle microgrid energy management, considering the volatility of the demand, renewable energy availability, real-time grid tariffs, and power flow constraints. Authors in [24] explored the feasibility of applying RL to schedule energy in a grid-connected PV-battery electric vehicle (EV) charging station. From the results, the algorithm managed to successfully obtain a day-to-day energy schedule that decreases the transactive cost between the microgrid and the utility grid. Authors in [25,26] proposed a battery management strategy in microgrids using RL technique. However, the incorporation of the battery wear cost in the EMS model was absent. The work in [27] used RL to develop a real-time incentive-based demand response program; the RL algorithm focused at aiding the service provider to buy power from its subscribed customers to balance load demand and power supply and improve grid reliability. Lu et al. [28] leveraged RL to design a dynamic pricing demand response (DR) algorithm in a hierarchical electricity market. From the results, the algorithm is seen to successfully balance energy supply and demand and reduce energy cost for consumers. Nakabi and Toivanen [29] proposed a new microgrid architecture consisting of a wind generator, an energy storage system (ESS), a collection of thermostatically controlled and price-responsive loads, and a utility grid connection. The proposed EMS was modelled to coordinate the different energy sources. Different scenarios were investigated using various deep RL methods. The proposed A3C++ algorithm was established to have an improved convergence and it also acquired superior control policies. In [30] a microgrid control problem focusing on energy trading with the utility is formulated. A deep Q-learning algorithm is used to learn the optimal decision-making policies. Simulation results, on real data, confirmed that the approach was effective, and it outperformed the rule-based heuristics methods. Samadi, et al. [31] proposed a multi-agent based decentralized energy management approach in a grid-connected microgrid. The different microgrid components were designed as autonomous agents who adopted model-free RL approach to optimize their behavior. Simulation results confirmed that the proposed approach was efficacious. Shang, et al. [32] proposed an EMS model aimed at minimizing the microgrid's operation cost, considering the nonconvex battery degradation cost. A RL method combined with Monte-Carlo Tree Search and knowledge rules is used to optimize the system. Although, the simulation results show the efficacy of the proposed algorithm, a detailed model of battery degradation is not considered in [32].

In recent advances reported on the implementation of RL in microgrid energy management [20–23,25–31,33–35] modelling of microgrid operational cost with consideration of battery degradation cost is not yet thoroughly studied. Most studies only consider the generation cost and power exchange cost. The estimation of the degradation process is very difficult and finding a simple and precise mathematical degradation model that can be used in the energy management algorithm is not easy. As the charging and discharging behaviors of a BESS have a direct impact on its life span, lifecycle degradation costs should be factored into the complex dispatch model of BESSs [36]. It is important to note that

Lithium ion batteries are quite expensive and incorporating a battery degradation model while computing the overall system cost is critical as a realistic system cost estimate is established. Thus, this paper reports on the development of an EMS for a grid-tied solar PV-battery microgrid considering battery degradation in the energy trading process, with the focus on reducing the strain on the battery. The aim of the designed EMS is to manage energy flows from and to the main grid by scheduling the battery such that the overall system cost (including cost of power purchased from the utility and battery wear cost) is reduced, and utilization of solar PV is maximized. The EMS problem is modelled as a Markov Decision Process (MDP) that fully explains the state set, action set and reward function formulation. In addition, two case studies have been considered where, in the first case, energy trading with the utility grid is permitted, whereas in the second case, it's not. To minimize the operational costs, a Q-learning based algorithm is implemented to learn the control actions for battery energy storage system (BESS) under very complex environment (e.g., battery degradation, intermittent renewable energy supply and grid tariff uncertainty). Simulation results show that agent learns to improve battery actions at every time step by experiencing the environment modelled as an MDP.

The key contributions of this work are outlined below:

Considering the technical constraints of the BESS, and the uncertainty of solar PV generation, load consumption, and grid tariff.

Developing an EMS architecture for a grid-tied solar PV-battery microgrid and formulating the control problem as a MDP considering the state, action, and reward function. The investigation of incorporating microgrid's constraints such that no power is scheduled back to the utility is also presented.

Using RL algorithm to learn the electrical resources and demand patterns such that system costs are reduced, and an optimized battery schedule is achieved.

Simulations results verify that the proposed algorithms substantially reduce daily operating costs under typical load demand and PV (summer and winter) generation data sets.

The novelty of the paper is presenting the design of an energy storage strategy that focuses on energy consumption optimization by maximizing the use of available PV energy and energy stored in the battery instead of focusing solely on direct storage control. In this architecture excess microgrid energy can be sold back to the utility to increase revenue however a non-trading algorithm scheme has also been studied where constraining rules are embedded into the learning process to curtail excess energy from been sold back to the utility. In addition, a battery degradation model is incorporated to reduce strain on the battery during the (dis)charge operation.

The rest of the paper is structured as: Section 2 presents the EMS problem formulation and introduces the two costs models considered i.e., grid transaction cost and battery degradation costs. Section 3 presents the MDP framework for the EMS problem formulation. Section 4 explains the proposed Q-learning algorithm. Section 5 presents the simulation setup, Section 6: Results are presented, and the algorithms performance are evaluated, and Section 7: recaps the paper's major points and introduces future work ideas.

2. Energy Management System Problem Formulation

This section presents a brief description of the EMS and then presents the MDP framework. This work considers a microgrid that consists of a PV system, a group of batteries, and some local loads as illustrated in Figure 1. The microgrid is capable of exchanging energy with the main grid at rates set by the utility company. Time-of-Use (ToU) grid tariffs have been adopted. Energy produced by the solar PV is used to meet the load demand at the beginning of every time step and is denoted by P_t^{PV} (kW). Excess energy produced by the PV during low energy demand can charge the battery. The battery has a maximum capacity denoted as E (kWh). It is also presumed that there are no charge and discharge losses. The microgrid system has an EMS for scheduling power flows to and from the main grid and manage battery charge and discharge.

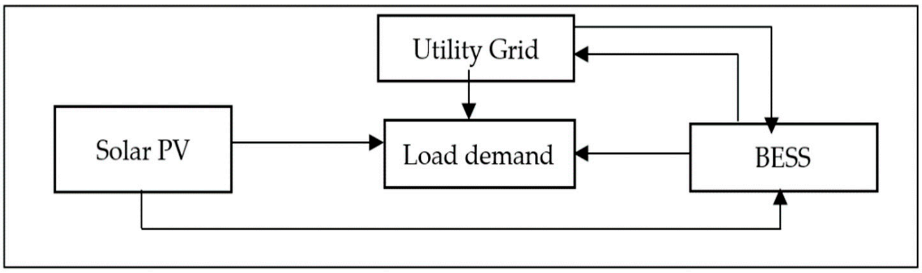


Figure 1. Illustration of grid-tied solar PV-battery microgrid.

2.1. Objective Function

A real-world microgrid system seeks to supply its total load demand using minimal energy cost. On this basis, the objective function of the designed grid-tied microgrid is computed as (1).

$$min\left\{\sum_{t\,=\,1}^{t\,=\,24}[(C_{\ g}(t)+C_{deg}(t))]\right. \tag{1}$$

Equation (1) defines the need to minimize the daily energy cost (i.e., over a 24-h horizon); $C_g(t)$ is the cost associated with grid and $C_{deg}(t)$ is the battery wear cost (the two cost components are expressed in R/kWh). The two cost models that the EMS tries to optimize will be illustrated below. Then the mathematical model of the EMS and all system constraints will be explained in Section 3 that presents the MDP framework.

2.2. Battery Degradation Cost Model

To formulate the battery's wear cost, stress factors that affect battery life are considered. In general, there are two forms of battery ageing: (i) calendar ageing and (ii) cyclic ageing [36]. The battery's intrinsic deterioration over time, which is influenced by temperature and an excessively high or low state of charge, is reflected in calendar aging. Whereas the capacity lost each time the battery is charged and discharged is referred to as cyclic aging, and it is dependent on the depth of charge, discharge rate, ambient temperature, and other factors. Excessively high or low state of charge (SoC) will degrade battery charging and discharging efficiency significantly. However, to avoid over-charging or over-discharging, the battery's SoC can be kept within a reasonable range by constraining the maximum and minimum SoC as seen in Section 3.1. Temperature can also have a detrimental effect on battery life, as the decay process is accelerated at high temperatures. In practice though, the temperature controller is often used in the battery management system. As a result, it is presumed that battery degradation caused by thermal heating factors can be ignored. Since DoD related stress have a significant impact to battery cycle life and represent a proper estimation of battery degradation, this paper will only consider the effect of depth of discharge on the battery [37].

DoD is described as a function of the battery's SoC and is depicted as [38] $DoD(t) = 1 - SoC(t)$. Authors in [39] researched on the relationship between lithium ion battery DoD and its life cycle data and established that the battery's life cycles increase exponentially with a reduction in the DoD as,

$$L(DoD)=\ \alpha DoD^{-\beta} \tag{2}$$

In (2) α and β are the curve fitting constants and the authors discovered that they are 694 and 0.795 respectively. The battery wear cost (C_{deg}) resulting from related dynamics of the battery life-cycle depicted by (2) considering a battery that operates from DoD_1 to DoD_2 with $DoD_2 > DoD_1$, could be approximated with (3) as shown below [38],

$$C_{DoD}=\ C_{bt}\left|\left(\frac{1}{L(DoD_{\ 2})}-\frac{1}{L(DoD_{\ 1})}\right)\right| \tag{3}$$

$L(DoD_j)$ denotes the battery's life cycle at DoD_j computed by (3) and C_{bt} is the initial capital investment of the battery per kWh. The cost of degradation is assumed to be independent of the direction of power flow in the battery, hence absolute values are considered by the solver. Finally, the battery degradation costs of the control action are calculated as,

$$C_{deg}^{t}= -(C_{DoD}(t)\Delta p(t)\Delta t\ ) \tag{4}$$

where, Δp is measured in kW and Δt in hours.

2.3. Utility Grid Model

The main grid can have two states: ON (available) and OFF (unavailable) and can supply the unmet load demand or/and charge the battery adequately for the microgrid whenever it is in ON state. At a given time step Δt, the microgrid must either be supplying power to the grid network through the battery system or buying power from the grid system (but not both actions at the same time) through the point of common coupling (PCC). Let $G_t(t)$ denotes the instantaneous grid tariff given in (R/kWh). In most cases, the selling price is usually lower than the purchasing price in order to encourage local use of solar PV power and minimize negative effects of microgrid uncertainty on the utility grid [40]. The microgrid selling rates are modelled as a discounted factor ϑ of the ToU tariff. Thus, the cost of exchanging energy in the microgrid is enumerated as,

$$C_g(t)= -(G_t(t)P_{g_p}(t).\ \Delta t - \vartheta\ G_t(t)P_{g\_s}(t).\Delta t\ ) \tag{5}$$

where $0 < \vartheta < 1$, $P_{g\_p}(t)$ denotes power purchased from the main grid and $P_{g\_s}(t)$ depicts the power sold to the utility system at every time step Δt. This power is elaborated in detail in (12) in the next section. If $C_g(t)$ is negative it indicates a cost to the microgrid as power is being purchased from the main grid, and when positive it depicts the profits gained from the microgrid selling power to the utility. The instantaneous grid power constraints are set as $0 \le P_{g_p}(t) \le P_{g_p}^{max}, 0 \le P_{g_s}(t) \le P_{g_s}^{max}\ \forall t \in t$, and $P_{g_p}(t) \cdot P_{g_s}(t) = 0$. The microgrid owner and the distribution system operator (DSO) have a contract that governs the maximum power that can be exchanged between the microgrid and the utility at the point of common coupling (PCC).

3. Markov Decision Framework as Applied to EMS Formulation

Markov decision framework or MDP is a mathematical framework used to model decision-making in situations where results are partly random and partly controllable and has been broadly adopted to map optimization problems solved through RL [41]. An MDP is defined as a four-tuple (S, A, T, R), where S and A are the state and action space, T and R denote the state transition probability, and the reward function respectively. Since, for this case the state transitions are deterministic, state transition modelling is not necessary [42] and only the state space, action space, and reward function are considered.

3.1. State and State Space Formulation

The information provided by the state is essential for energy management as it contains the information that the agent uses in the decision-making process at each time step t. The state space of the EMS at any given time is defined by the utility tariff (R/kWh), the BESS state of charge, the load demand (kW) and the PV generation (kW).

Let the state of charge of the battery at time step t be denoted as $SoC = \{SoC_t\}$.

So as not to exceed the battery constraints, a guard ratio β is considered as, $\beta \cdot E \le SoC_t \le (1-\beta) \cdot E$, where $\beta \in [0, 0.5]$ [19] and E denotes the energy capacity of the battery (kWh).

At each time the state of charge of the battery is constrained by, $SoC^{min} \le SoC_t \le SoC^{max}$, where SoC^{min} and SoC^{max} represents the lower and the upper bounds of the battery.

Considering the above battery safety limits, the state s_t at each time step t is,

$$s_t \ = \ \left\{ t, P_t^{PV}, SoC_t, G_t, P_{l,t} \right\} \tag{6}$$

where t is the time component denoting the hour of the day, P_t^{PV} is the generation from solar PV at time t, G_t denotes the current electricity tariff at time t notified by the utility company, $P_{l,t}$ is the instantaneous load demand. The state space is enumerated by the union of all set of states within the optimization horizon as, $S \ = \ s_0 \cup \ s_1 \cup, \ldots \cup s_{T-1}$. The intraday microgrid operation has been divided into T timesteps, indexed as {0,1, 2 . . . , T−1}, where T represents the optimization horizon under consideration.

3.2. Action and Action Space Formulation

In order to meet the load demand in every time step Δt, the EMS of the microgrid first uses the available energy from the solar PV and the BESS, then the remaining energy is purchased from the utility. Net load $P_{l,t}^{Net}$ of the microgrid at each time step t is described as the total demand $(P_{\,l,t})$ minus the energy generated by the solar PV $\left(P_t^{PV}\right)$ as shown below:

$$P_{l,t}^{Net} = max\ \left((P_{\,l,t} - \ P_t^{PV}),\ 0 \right) \tag{7}$$

Here, "*max*" ensures that the complier takes the maximum value always. For instances, if the PV is large than the load, that equation will output a negative value, which is not the case as the net load is not negative. To prevent that a zero is put (it will be the max value at that time step) meaning the load has fully been covered by the solar PV.

Since the total load demand $P_{l,t}$ and PV generation P_t^{PV} fluctuate stochastically in a real microgrid, the net demand of the microgrid, $P_{l,t}^{Net}$ is an unknown variable. First, the EMS tries to satisfy the net demand $P_{l,t}^{Net}$ through the energy stored in the BESS. Then, the remaining load demand that cannot be covered by the BESS is provided by the utility. It is described as the reminder energy $P_{l,t}^{rem}$ which can be enumerated as:

$$P_{l,t}^{rem} = max(P_{\,l,t}^{\,Net} - \left(SoC_t - SoC^{min} \right).E,\ 0) \tag{8}$$

The amount of energy that need to be purchased at each time step is denoted as $P_{l,t}^{rem}$. At each time step, after covering the load demand the quantity of energy contained in the BESS denoted as SoC_t^{next}, is calculated as shown in (9).

$$SoC_t^{next} \ = \ min(SoC^{\,max}, (max\left(\ P_t^{PV} - P_{l,t}\ ,\ 0 \right) + max\left(\frac{\left(SoC_t - SoC^{min} \right) \cdot E - P_{l,t}^{Net}}{E}, 0 \right)) \tag{9}$$

This equation is generally computing the amount of energy remaining in the battery. The first section checks if there is any remaining solar power after supply the load, if yes, the solver will charge the battery, if there isn't, zero will be taken. Since the EMS is designed to first check if there is any energy in the battery before purchasing from the utility as show in (8), the second part of the equation calculates the remaining energy in the battery after supplying the load so that we can have the accurate state of charge for the next time step.

Since the agent can only dispatch the battery, i.e., manage charge and discharge. To simplify this problem, the actions are discretized here into discharging/charging action category. The power unit Δp depicts the amount of power that is used to discharge/charge the battery in each discrete instant. The discrete action space is defined as

$$A_{s_t} \ = \ \{ -k\Delta p, \ldots, -\Delta p, 0, \Delta p, \ldots, k\Delta\, p \ \}, \tag{10}$$

where $k\Delta p$ and $-k\Delta p$ are the maximum amount of charge and discharge power from the BESS in each time step respectively, while 0 indicates that the battery is idle. $a_t \in A_{s_t}$ is defined as the action selected at time step t by agent, where A_{s_t} represents all the possible actions in the action space A under state S_t.

Given the action set A_{s_t} in Equation (10), at every time step Δt, the agent chooses one possible a_t, from A_{s_t} by following a policy π, that describes a decision-making strategy for the selection of actions. More details on π can be found in the next section.

Let the function of the amount of power supplied to the battery when an action a_t, is taken by the agent be denoted as BESS (a_t) and computed as,

$$\text{BESS}(a_t) \ = \left\{ \begin{array}{c} -\frac{k(a_{\ t})}{E} k(a_{\ t})/E, \text{ if } a_t = \text{discharging} \\ \frac{k(a_{\ t})}{E}, \text{ if } a_t = \text{charging} \end{array} \right\} \tag{11}$$

where the negative values indicate discharge from the battery and positive values indicate charging of the battery. The result of the agent action BESS(a_t) to the battery is based on the status of the BESS SoC_t^{next}.

It is presumed that if the action taken a_t(charging) increase the $SoC_t + k(a_{\ t})/E$ past the maximum guard capacity E^{max}, only the energy chargeable $SoC^{max} - SoC_t$ is used to charge the battery and the extra energy is discarded. Similarly, for the discharging action, only $SoC_t - SoC^{min}$ is discharged and the extra discharge energy is discarded, hence the battery constraints are never violated.

3.3. Reward Function Formulation

Reward is a scalar value used to express to the agent the goal of the learning process. Once the agent performs an action and moves to the next state, a reward is presented. Intelligent "reward engineering" is key as it links the agent actions to the objective of the algorithm [43]. The objective of the optimization process is to minimize the transaction cost of power purchased from the utility and reduce battery wear cost.

Reward $r(s_{\ t}, a_t)$ of the proposed EMS is structured to evaluate two aspects of the system management, one is the objective function and the other two aspects suggested by [44] are adopted to improve the agent's performance. The objective function factors in the amount of money incurred by purchasing energy from the main grid C_g, and battery degradation costs C_{deg}. To improve algorithm performance, C_b and C_o have been incorporated. C_b represents gains from pre-charged energy and C_o is a penalty payment charged to the agent when it chooses an action that exceeds the limits of the battery.

The pay reward C_g^t represents the cost incurred by trading power with the utility at each time step. The agent receives a negative reward if the amount of energy purchased from the grid is greater than the amount of energy sold. Otherwise, the agent will receive a positive reward of C_g^t calculated as given below.

$$C_g^t = -\left(P_{l,t}^{rem} + BESS(a_t).E\right).G_t \tag{12}$$

In (12) $P_{l,t}^{rem}$ represents the total unmet load in the microgrid at each time step (kWh) while G_t denotes the instantaneous grid tariff (R/kWh). The sum total of $P_{l,t}^{rem} + BESS(a_t).E$ indicates the power being exchanged with the utility grid at each time step Δt.

In the non-trading mode of operation, the energy supplied to the load (when a discharge action is selected) at any time slot cannot be higher than the load demand. Equation (13) ensures energy cannot be sold back to the utility. During training if the learning agent tries to select actions that causes power to being scheduled back to the utility a small negative penalty C_p^t will be charged.

$$P_{l,t}^{rem} + BESS\left(a_{t.discharging}\right).E \geq 0 \tag{13}$$

$$C_p^t = P_{l,t}^{rem} + BESS\left(a_{t.discharging}\right).E.G_t.\vartheta \tag{14}$$

Next, C_b^t is computed as the amount of available energy in the battery to cover the net load demand $P_{l,t}^{Net}$ from the energy stored in the BESS SoC_t. This reward mainly encourages the agent to always ensure that the SoC of the battery can satisfy the net load at any time. When current grid tariff G_t increase, this benefit reward increases as well. In simple terms, the reward reflects reduced payment that result from using the battery instead of purchasing power from the grid.

$$C_b^t = \begin{cases} P_{l,t}^{Net} \cdot G_t & if\ P_{l,t}^{Net} \le \left(SoC_t - SoC^{min}\right).E \\ \left(SoC_t - SoC^{min}\right).G_t.E, & else \end{cases} \tag{15}$$

Then, C_o^t as shown in (14) below, represents a penalty received by the agent at each time step for any extra energy supplied but is not used in the charging/discharging of the battery due to enforced constraints. As the grid tariffs G_t increases, the over-charged penalty becomes high.

$$C_o^t = \begin{cases} -((SoC_t + k(a_t) - SoC^{max}).G_t.E & if(SoC_t + k(a_t) > SoC^{max} \\ -(|k(a_t)| - (SoC_t + SoC^{max})).G_t.E & elif(SoC_t + k(a_t) < SoC^{min} \\ 0 \quad else \end{cases} \tag{16}$$

Finally, the cost of battery degradation C_{deg}^t is considered as a negative reward received by the agent and it is calculated as show in (4).

Let $r(s_t, a_t s_{t+1})$ denote the cumulative reward that the agent receives when it takes an action a_t at state s_t. The total reward that the agent gets at each time step is given by (17), however in the non-trading mode of operation C_p^t is incorporated in Equation (17)

$$r(s_t, a_t) = C_g^t + C_{deg}^t + C_b^t + C_o^t \tag{17}$$

As an RL agent traverses the state space, it observes a state s_t takes an action a_t and moves to the next state, s_{t+1}. In order to compute the impact of an action taken by the agent on future rewards while following a certain policy π, $V_t^{\pi}(s)$ has to be computed. It is defined as the cumulative discounted rewards at time slot t and calculated as

$$V_t^{\pi}(s) = r(s_t, a_t) + \sum_{i=1}^{\infty} \gamma^i r(s_{t+1}, a_{t+1}) \tag{18}$$

The first term in (18) is the immediate reward at time step t and the second term is the discounted rewards from the next state s_{t+1}. Here, $\gamma \in [0,1]$ is the discount factor, which determines the weight given to future rewards by the agent, where a high value makes the agent more forward thinking. π is used to represent a stochastic policy that maps states to actions: $\pi(s_t, a_t) \rightarrow \mathrm{S} \times A$. The agent's goal is to find a policy π (battery schedules) that maximizes the long-term discounted rewards. An optimal policy π^* is the MDP's solution, i.e., a policy that constantly selects actions that maximize the cumulative rewards for the (T) hours horizon starting from the initial state s_0 [14]. To solve the MDP, several RL techniques can be applied. Model-based methods, such as Dynamic programming (DP), assume that the dynamics of the MDP are known (i.e., all state transition probabilities). On the other hand, model-free techniques such as Q learning learn directly from experience and do not assume any knowledge of the environment's dynamics. To get the solution of the MDP designed above Q-Learning has been adopted and it is explained in detailed below.

4. Q-Learning Algorithm for Energy Management Problem

Q-learning is the most widely used model-free RL algorithm i.e., it can implicitly learn an optimal policy (a sequence of battery action selection strategy) by interacting with the environment without any prior knowledge of the environment (as opposed to model based methods where the agent has to learn the entire dynamics of the system then plan to obtain the optimal policy) [14]. Q-learning involves the finding of the so-called Q-values where Q-values are defined for all state action pairs, (s, a). The Q-value gives the measure of goodness of selecting an action a in state s.

Let $Q(s, a)$ represent the State-Action value function that computes the estimated total discounted rewards as calculated in (20), if an action a_t is executed at state s_t when a policy π is followed. It will be described as,

$$Q(s_t, a_t) = E\{V_t^{\pi}(s)\} \tag{19}$$

$$Q(s_t, a_t) = E\{r(s_t, a_t) + \sum_{\mathrm{i}=1}^{\infty} \gamma^i r(s_{t+1}, a_{t+1})\} \tag{20}$$

where $\mathbb{E}$ indicates the expected action value for each state action pair.

The Q-value that reflects the optimal policy is denoted as $Q^*(s,a) = Q^{\pi^*}(s,a)$, $\forall s \in S$, $\forall\, a \in A_{s_t}$. If all possible actions in each state s are selected and executed multiple times in the environment and their Q-values updated a sufficient number of times, then Q-values eventually converge [16] and the optimal action in that state can be found by taking the action that maximizes the Q-values. The optimal Q-value is given by,

$$Q^*(s,a) \ = \ \begin{matrix} max \\ a \end{matrix} \ Q^{\pi}(s,a), \forall s \in S, \forall a \in A_{s_t} \tag{21}$$

And the optimal policy is acquired as (22) for each state s,

$$\pi^*(s) = argmax_{a \in A} Q^*(s,a) \tag{22}$$

Equation (22) implies that an optimal action-value in any state s is described as $Q^*(s,a^*) > Q^*(s,a_i)$, $\forall a_i \neq a^*$, where a^* is the optimal action for state s, commonly known as the greedy action a_g. During the learning process, the agent interacts directly with the dynamic environment by performing actions. Generally, the agent observes a state s_t as it occurs, with the possible action set A_{s_t}, and by use of an action selection technique, it selects an action a_t and consequently, moves to the next state s_{t+1}, and receives an immediate reward, $r(s_{\,t}, a_t, s_{t+1})$. Then updating of the Q-values is done based on the Bellman equation as shown in (23),

$$Q^{n+1}(s,a) = Q^n(s,a) + \alpha\left[\, r(s_t, a_t, s_{t+1}) + \gamma max_{a_{t+1}} Q^n(s_{t+1}, a_{t+1}) - Q^n(s,a)\right] \tag{23}$$

where $\alpha \in$[0,1] denotes the learning rate which determines the extent by which the new Q-value is modified, $Q^n(s,a)$ is the current estimate of Q-value, $Q^{n+1}(s,a)$ represents the next estimated Q-value in the next iteration, whereas $\gamma \in [0,1]$ denotes the discounting factor and n is the specific iteration number. When α is sufficient small, and all possible state-action pairs are visited enough times Q^n eventually converges to the optimal value Q^* so that best action will be selected at each state in the successive iterations [16]. When the agent reaches the terminal state s_{T-1}, since there are no future rewards, the Q-value is update as shown in (24) below:

$$Q^{n+1}(s,a) = Q^n(s,a) + \alpha\left[\, r(s_t, a_t, s_{t+1}) - Q^n(s,a)\right] \tag{24}$$

As an agent chooses actions from the action set, it is always necessary to cleverly deal with the exploitation versus exploration dilemma [11,45]. Exploration helps the agent to avoid getting stuck in a local optimum while as exploitation allows the agent to selected best actions in the later episodes. Epsilon greedy (ε_{greedy}) method is adopted here because of its simplicity. Epsilon greedy is a method of selecting actions with uniform distribution from an action space. Using this strategy, it is possible to select a random action (exploration) from the action space A_{s_t} with probability ε. It is also possible to choose a greedy action (exploitation) with probability $1 - \varepsilon$, for $\varepsilon \in [0,1]$, from the Q-values at the given state in each episode. An exponential decay function is also leveraged, so in each iteration, the value of ε is modified as follows; $\varepsilon = \varepsilon_{min} + (\varepsilon_{max} - \varepsilon_{min}) exp\{-\mathtt{C} \times n\}$, where ε_{min} and ε_{max} represents the minimum and maximum values of ε respectively, $\mathtt{C}$ is the exponential decay rate and n denotes total number of iterations.

It is to be noted that epsilon ε varies from case to case depending on system design. But the idea is to allow the agent to explore all the actions in the initial episodes so as to learn. As learning proceeds ε epsilon should gradually be decreased to enable the agent to choose greedy actions. But we should still leave a very small percentage for taking a random action as there is a probability that current estimate may be wrong and there is another better action. For practical problems during training start with a very large number of epsilons i.e., ε = 1 and keep lowering that value to 0.001 or 0.01. so that the agent can exploit the best action in the final iterations.

Algorithm for Learning Energy Management

To tackle the MDP, a Q-table is first created and initialized with zeros. At the beginning of the learning, initialization of hyperparameters γ, $\propto$, and ε is done in lines 2–3 of the algorithm shown below. Lines 6 to11 shows the loop for every time step Δt. In line 5 the microgrid environment is initialized, while in line 6, the algorithm reads the current state. In line 7 action a_t, is selected depending on the action selection policy π. In line 8, the selected action is executed in the environment and the environment produces a reward $r(s_{t,} a_t)$ and the next state s_{t+1}. Based on the return of the environment, $Q(s_t, a_t)$ is updated according to Equation (19) and if it's a terminal state update is done by (20) in line 9; In line 10, the time step t is incremented by one, $t + 1$ and the system move to the next state. After the terminal state T$-$1, the next episode proceeds with an updated value of ε. Then, the learning process continues as seen in Algorithm 1 below.

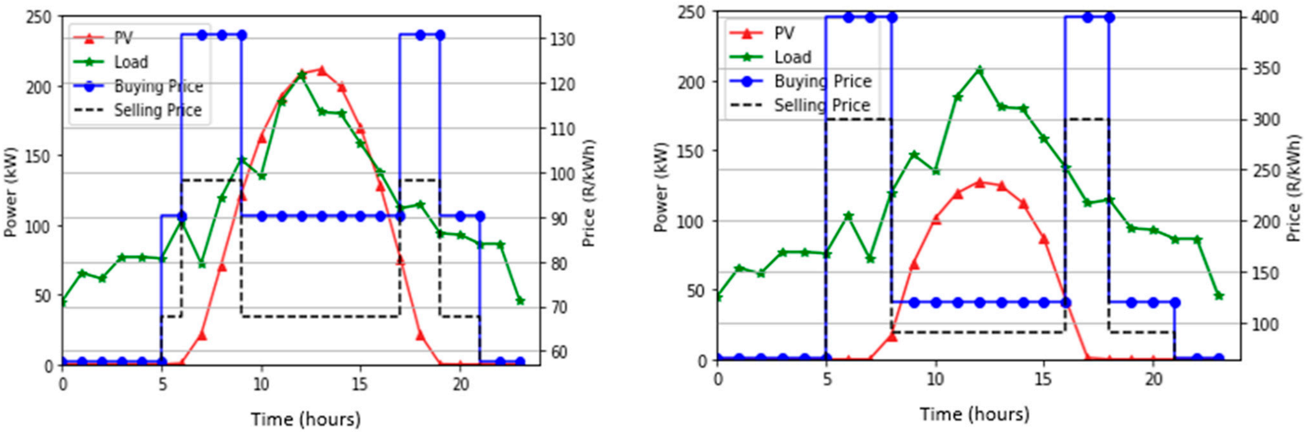
Algorithm 1 EMS Algorithm Using Q Learning

```
1. Create a q-table and initialize $Q(s,a) \forall s \in S, \forall a \in A$, with zeros,
2. Initialize learning rate and gamma ($\propto$ and $\gamma$)
3. Initialize epsilon ($\varepsilon$)
4. For episode ($n$) = 1, $max\ Episode$ do
5.   Initialize Microgrid Environment
6.   For time step ($t$) = 0, T$-$1 do
7.     Read the current state
8.     Select an action at using $a_t$ from $A_{s_t}$ using the $\varepsilon$ greedy policy $\pi^{\varepsilon}(s)$ (5)
9.     Execute the selected action $a_t$ in the Simulation Environment and observe the reward $r_t$ and the
       next state $s_{t+1}$
10.    Update q-values according to (19)
11.    $t = t + 1$
12.  End
13.  Update $\varepsilon$
14.  $n = n + 1$
15. End
```

5. Simulation Setup

To evaluate the performance of the proposed energy management algorithm using Q-learning, this work considers a commercial load grid-tied microgrid environment with solar PV and BESS. Numerical simulations are performed based on a commercial building load profile data adopted from [46]. Summer and winter solar PV output data for (November (summer) and June (winter)) in a 250 kWp solar PV system located in Cape Town, South Africa adopted from [47] are used in the simulation. To facilitate the assessment of optimized control strategy the work considers an hourly time of use (ToU) tariff obtained from Eskom a utility company operating in South Africa which specifies three price levels applied based on time of day during summer and winter seasons. Peak prices are equivalent 130.69 R/kWh, mid-peak prices equal to 90.19 R/kWh and off-peak prices equal to

57.49 R/kWh during summer while during winter peak tariff is 399.17 R/kWh, mid peak is 121.46 R/kWh and off peak price is 66.27 R/kWh [48]. The forecasted time series inputs to the algorithm which include the commercial load demand and solar PV generation are shown in Figure 2a,b for summer and winter season respectively. The peak load of the commercial consumption profile is noted to occur between 09:00 and 16:00 when most HVAC and loads are switched on. For the BESS, two Lithium-ion batteries are used, where each battery has a capacity of 200 kWh. The initial SoC of the BESS is set to 0.25, and the guard ratio $\beta = 0.05$ is considered since any value in this range $[0, 0.5]$ can be selected. Thus, the maximum and minimum limit of the BESS are set up to $E^{max} = 380$ kWh and $E^{min} = 20$ kWh respectively. The initial battery cost is determined based on the current market price of Li-ion battery which is 135 USD/kWh (2025 R/kWh) [49]. The charge and discharge power unit Δp is set to 25 kW, where the charge power of BESS is uniformly discretized to k is 6. Thus, the discretized charging and discharging power of the battery is, $A = \{-150, \ldots, -50, -25, 0, 25, 50, \ldots 150\}$ in kW, 150 and -150 represent the maximum charging and discharging power; 0 indicates the battery is idle, while the rest are values within the limit's interval. The maximum charge and discharge power are limited to 150 and -150 to ensure safe battery operation limit while the charge and discharge power unit is set to 25 kW to give the agent more variables in the action space. For simplicity purposes, power inverter efficiencies for Solar PV and battery is assumed to be 1. The algorithm is implemented in Python programming (version 3.7.6) and executed by a computer with a 1.60 GHz processor and 8 GB RAM.



(**a**)

(**b**)

Figure 2. Input data; (**a**) Summer solar PV power, load profile, and summer grid prices. (**b**) Winter solar PV power, load profile, and winter grid prices.

It is critical to properly select parameters, especially those to which the algorithm is highly sensitive, such as the learning rate and the discount factor, in order to achieve a suitable convergence speed and quality policies. If a large step-size rate is selected, $Q(s, a)$ values can oscillate significantly and if it is too small, Q-values might take long before they converge. The choice of α was by trial and error and a value of 0.01 gave the best convergence. The ε-greedy parameter ε was initialized to 1 to ensure the entire search space is explored as much as possible, and a discount factor γ of 0.85 (for winter case) and 1 (for summer) is taken as the future rewards are significantly important as the immediate rewards. The simulation input parameters for the EMS algorithm can be seen in Table 1.

In order to evaluate the performance of the proposed grid-tied microgrid energy management system, two case studies are simulated on the basis of the data characteristics mentioned above. First, two different seasons are examined to assess the impact of PV penetration. Second, the comparison between including and excluding grid constraints at the interconnection point is then performed with the aim of studying the impact on total operating costs. In the case of grid constraints (non-trading algorithm) Equations (13) and

(14) are included in the optimization model to ensure that the microgrid does not sell its surplus energy back to the utility grid, while for no-grid constraints (trading algorithm) they are removed.

Table 1. Simulation Parameters.

Hyperparameters	Selected Values	Values
Epsilon	ε	1.0
Learning rate	α	0.01
Discount factor	γ	1 summer data/0.85 winter data
Timestep	Δt	1 h
Battery initial cost	C_{bt}	2025 R/kWh
Battery capacity	E_b	400 kWh
Initial SoC of the ESS	SOC_0	0.25
Battery guard ratio	β	0.05
Power unit	Δp	25 kW
Selling price discount factor	ϑ	0.75
Grid power limits (trading algorithm)	$P_{g_s}^{max}/P_{g_p}^{max}$	−150/250 kW
Grid power limits (non-trading algorithm)	$P_{g_s}^{max}/P_{g_p}^{max}$	0.0/250 kW

6. Results and Discussion

6.1. Summer Solar PV and Grid Tariff Profile

The performance of the proposed energy management system in a one-day summer operation will be assessed in the current section. The summer PV profile and the summer grid tariffs are considered. Summer solar PV is considered to be the best-case study in the trading algorithm as it is more profitable to increase operating revenues by selling any excess energy back to the utility grid. The total produced energy by PV during summer is 1587 kWh.

6.1.1. Reward Convergence during Summer

The primary assessment explores how the system performance is improved by the EMS algorithm as the learning process progresses. Figure 3a,b displays the training curves for the trading and non-trading case studies respectively, which show the average Q-learning algorithm's cumulative reward profile for 20,000 training episodes. Between episode 0 and 5000, the agent is still in the initial stages of learning and the reward curve starts at a lower average value of −R 140,000 for the trading algorithm and −R 175,000 for the non-trading as can be observed in the Figure 3a,b below (here negative values for the reward indicates a cost to the microgrid as power is being purchased from the main grid). This is because initially, for both cases, value of ε is set to 1.0, i.e., every action has equal probability of being selected as the action space is still being explored on a trial-and-error basis by the learning agent. Later, as exploration rate decays, and the learning agent starts to exploit the best actions, it is seen that the training curves begin to rise and then they converge at higher value at about episode 7500 for the trading algorithm and 8000 for the non-trading one. Convergence is achieved because the agent begins to select better actions learned through the process of experiencing more state-action pairs. It can be observed that the non-trading reward curve reaches a high value of about 0 compared to the trading algorithm which only achieves −R 60,000. The reason for this is non-trading algorithm has an additional negative penalty on the reward formulation if the grid constraint is violated as show in Equation (14) which is not present in the trading algorithm. It can be concluded that both proposed energy management schemes are able to achieve optimized policies and Figures 4a and 5a show the selected battery actions of the optimal policy for both the trading and non-trading algorithm respectively.

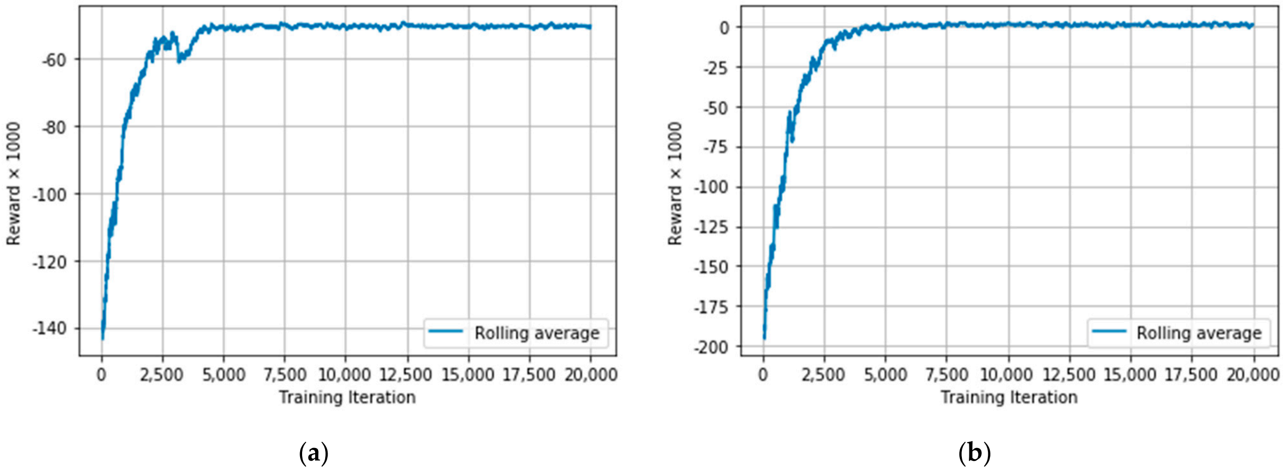


Figure 3. Training curve showing reward convergences for episode number 0 to 20,000; (**a**) trading algorithm (**b**) non-trading algorithm.

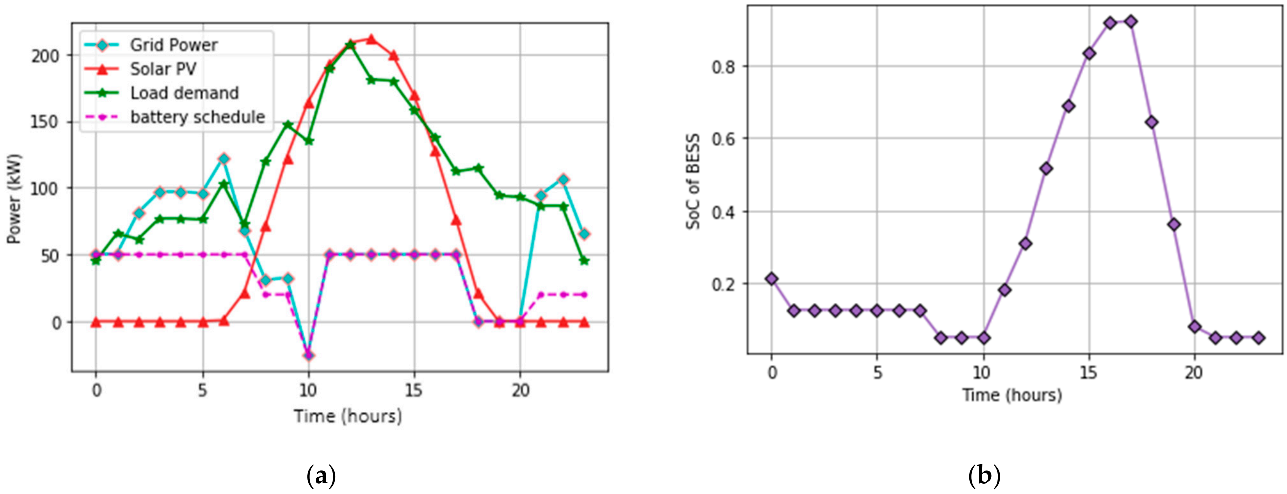


Figure 4. Observations for trading algorithm; (**a**) Optimized grid and battery schedules for 24 h horizon. (**b**) Tendency of the SoC of the BESS.

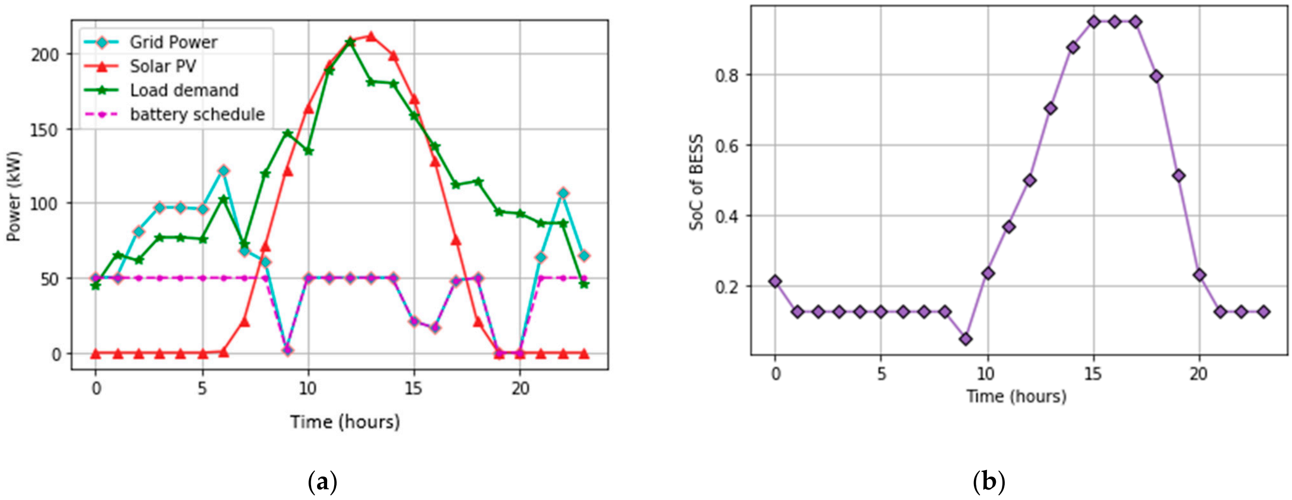


Figure 5. Observations with grid constraints; (**a**) Optimized energy schedule. (**b**) Tendency of the SoC of the BESS.

6.1.2. Energy Management

6.1.2.1. Results for Case Study 1 (Trading Algorithm)

This section presents the results of the trading algorithm which is executed through Equation (12) i.e., the agent obtains revenue by discharging the battery if the reminder power $P_{l,t}^{rem}$ is zero at any time step t. Analysis of how energy stored in the BESS is used as the EMS seeks to meet net demand is also carried out. When it comes to system running cost, charging the battery when tariffs are low and discharging the battery when tariffs are high is important so as to rip some revenue. Since the energy demand varies randomly, an efficient charging management algorithm should manage to effectively cope with any unanticipated event and still reduce system operational costs. Between 00:00 and 05:00 PV power is zero. Hence, in Figure 4b, a decrease in the SoC is seen since the battery is supplying the net load. Also, the load that is not met by the battery, however small, is met by the grid at low prices. Between 11:00 and 17:00, the SoC of the BESS is seen to gradually increase to 0.85 as the battery is being charged by the utility grid. The utility peak load occurs two times in a day i.e., between 07:00 to 09:00 and 18:00 to 19:00. During the first peak load, it is seen in Figure 4a that the algorithm learns to lower power intake from the utility to 25 kW. In the second peak we see the algorithm learns to raise the battery SoC to 0.85. From 17:00 to 20:00 battery SoC decreases because the battery is fully supplying the microgrid's net-load and zero grid power has been scheduled at that time as the prices are very high. From 21:00 to 23:00 a low SoC is seen as only 25 kW is being charged to the battery. A final SoC of 1.25 is recorded as seen in Figure 4b. Given the stochasticity of the load demand, grid tariff and solar PV, it is crucial that the battery energy can deal with unforeseen circumstances, and we can see the agent learns policies to increase the SoC to meet its load demand fully during peak tariff hours.

Figure 4a displays energy schedules of the grid and the battery plotted besides the solar PV and microgrid's load curve. Between 00:00 and 07:00 it is clearly seen that the algorithm opts to charge the battery with 50 kW throughout that period. From 02:00 to 06:00 a gradual increase in power absorbed from the utility is seen, because the battery cannot fully meet the net load, hence the unmet load is being covered by the grid. Furthermore, the tariff is very low (please refer to Figure 2a), and it would be optimal to utilize the cheap grid power to supply the net load and charge the battery. At 07:00 there is a sharp increase in grid tariff (R 40 increase is noticed), and the algorithm lowers the amount of power drawn from the utility by (25 kWh) for two consecutive hours. At 10:00, the agent sells 25 kW back to the utility during mid peak tariff thus maximizing its revenue. This would be evident by looking at Figure 4a and Figure 11a simultaneously. In Figure 4a, at 10:00 grid power is −25 kW and also in the same hour in Figure 11a, it is seen that R 2000 was deducted from total cost as power was sold to the utility at that hour. From 10:00 to 15:00 solar PV power is sufficient to fully cater for the load, however the algorithm opts to constantly charge the battery with 50 kW from 11:00 to 17:00. At 18:00, grid tariff shoots to its peak prices and it is clearly seen that the algorithm schedules zero grid power from 18:00 to 20:00 as the battery can fully cater for load even when solar PV is scarce. This shows that the algorithm manages to foresee the utility peak load and takes proactive decisions of buying power from the utility at mid-peak price, and thus shifts its load from 18:00 to 20:00. At 20:00 the grid prices decrease by R 40, and the agent beings to gradually increase utilization of the grid's power. A full utilization of the main grid is observed from 22:00 to 23:00 as the grid tariffs are at their lowest value and solar PV is not available at that time.

6.1.2.2. Results for Case Study 2 (Non-Trading Algorithm)

Figure 5a displays energy schedules of the grid power and the battery for the non-trading algorithm which constraints microgrid's power exchange with the utility such that at each time step no energy can be sold back to the grid as Equation (13) has been incorporated in the optimization model. In comparison to Figure 4a, it is seen that at 07:00 when the grid tariff is increased by R 40 (peak tariff) the algorithm doesn't reduce its battery power intake unlike in the trading case study. However, at 09:00 while the grid tariff is still peak, we see the algorithm schedules zero grid power, thus managing to support the grid by lowering its power intake for one hour. From 10:00 the algorithm raises the battery SoC by consecutively charging the battery with 50 kW and 25 kW. At 19:00 and 20:00, when solar PV is zero and grid tariff is high, it can be observed the agent shifts its load by scheduling zero power from the utility at that time. During off-peak prices at 21:00 and 23:00 we observe maximum usage of utility power as also power from solar PV isn't available. In Figure 4a the energy trading algorithm (case 1) is seen to sell 25 kW back to the main grid at 10:00am however, in case 2 in Figure 5a (and later in Figure 11b) where grid constraints are enforced no trading of power was observed. It can be concluded that both algorithms learn to reduce power absorbed from the main grid at utility's peak load demand during which buying prices are very high, however with the trading algorithm better policies are achieved as the operational cost is lower. Also, it learns to delay drawing power from the utility for 3 h (from 18:00 to 20:00) until the energy prices lower as seen in Figure 4a in contrast to the non-trading algorithm shown in Figure 5a that delays for 2 h (from 19:00 to 20:00).

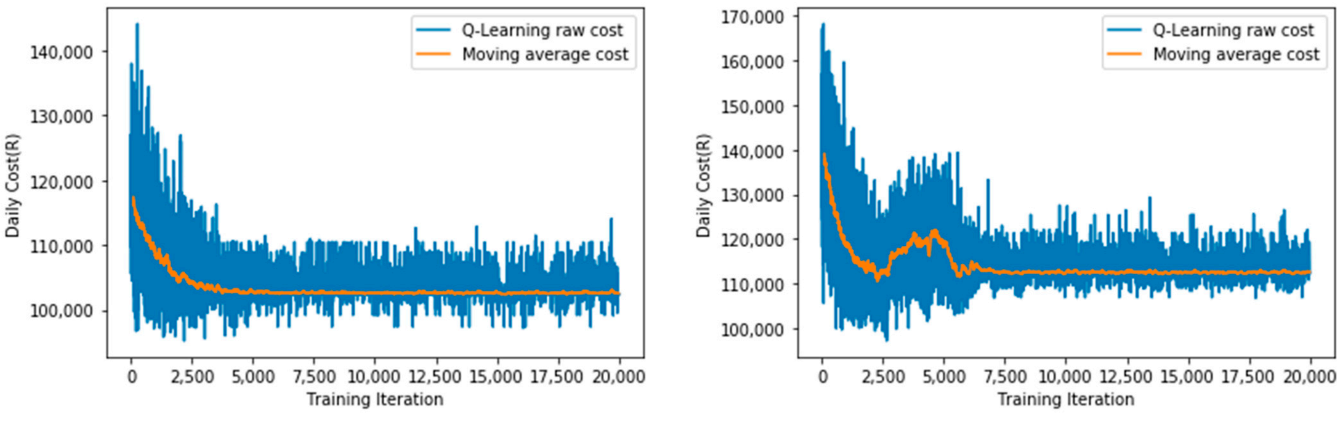
Figure 5b shows the battery SoC trajectories as the non-trading algorithm is being executed. Similar to Figure 4b from 01:00 to 09:00 the battery SoC decreases slightly and then remains constant since the battery is partly supplying the net load. Between 10:00 and 15:00, the peak load is catered fully by the PV, and the surplus solar PV can charge the battery. The second utility peak demand occurs between 18:00 and 19:00 when PV power is scarce, it can be observed that the SoC of the BESS gradually increases to 0.87 at around 15:00 to support the main grid during its peak demand. From 17:00 to 21:00 the SoC decreases as can be seen in Figure 5a zero power is scheduled from the utility for two consecutive hours and PV power is decreasing thus the battery is fully supplying the microgrid's net load. A final SoC of above 0.125 is recorded. The plot shows that learning agent learns to increase SoC to cope with any unanticipated uncertainties, maintains reasonable SoC trajectories throughout the 24-h horizon and ensure no battery's constraints are violated.

6.1.3. Operational Cost during Summer

Figure 6a,b represent the total daily operation cost plotted versus the training episode number. The moving average values are computed for every 100 episodes window. A decreasing trend can be noticed as the learning episodes increase. The daily operating cost at any time step is the grid trading cost and cost of battery degradation as shown by Equations (4) and (5). As can be seen in the graph, the agent explores different possible energy dispatches during the initial stages of learning and very high costs are registered during the initial stages of learning. For the trading algorithm an average value of about R 120,000 is registered and for the non-trading algorithm a value of is R 140,000 recorded. As the agent learns better policies, it begins to constantly exploit control actions which reduce energy cost in the final iterations. In the final episodes we can see the algorithm finishes at an average global cost of about R 105,000 for the trading algorithm and R 110,000 for the non-trading algorithm.

6.2. Winter Solar PV and Grid Tariff Profile

This section evaluates the behavior of the proposed EMS during a day operation in winter season. The winter PV profile is considered to be the worst-case study as the PV energy output is expected to be lower than summer output as a result of shorter daylight hours, change in the angle of the sun which reduces the sun's rays hitting solar panels, and extreme atmospheric conditions such as cloud covers and wet weather. The total energy output of PV production during the considered day amounts to a sum total of 801 kWh. Also, it can be noted that the winter tariff is rather high compared to the summer tariff as cold and dark weather cause people to stay indoors more, to turn on the lights for longer hours, and to switch on heating equipment, thereby increasing energy demand. In addition, extreme weather conditions could also damage the power system, resulting in high repair costs.

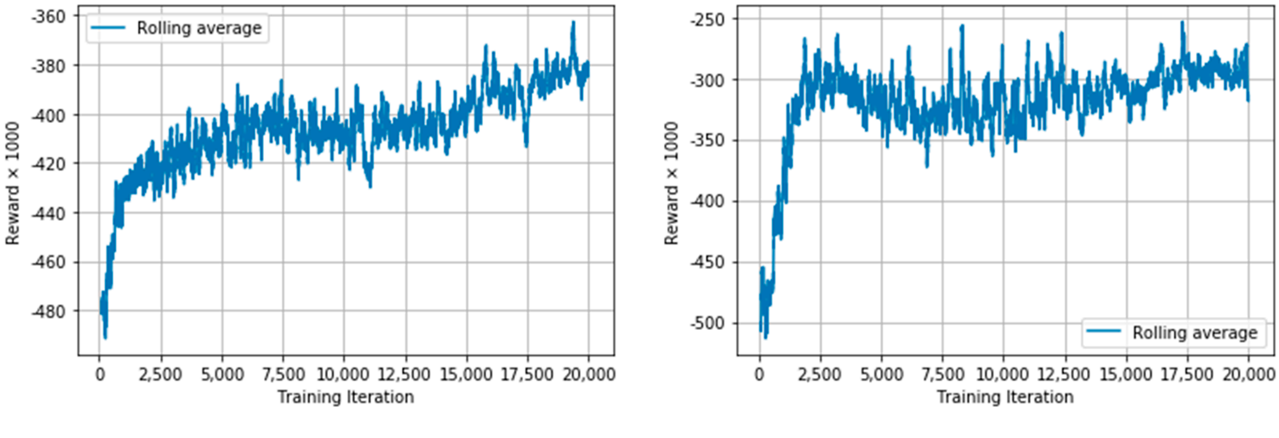


(**a**) (**b**)

Figure 6. Daily operational cost against number of episodes; (**a**) trading algorithm (**b**) non-trading algorithm.

6.2.1. Reward Convergence during Winter

In Figure 7a,b, it can be observed that both the trading algorithm and non-trading algorithm are capable of increasing the average reward over 20,000 training episodes. Between episode 0 and 5000, the agent is still in the initial stages of learning and the reward curve starts at a lower average value of −R 500,000 for the trading algorithm and −R 540,000 for the non-trading algorithm. This is because initially the learning agent is still exploring the stochastic environment on a trial and error. Later, as exploration rate decays, the learning agent starts to exploit the best actions, it is seen that the training curves begin to rise and then converge to higher values ats episode 12,500. It can be observed that the trading algorithm converges to a lower average value (−R 380,000) in comparison to the non-trading algorithm which converges at an average value of about −R 300,000. The reason for this is non-trading algorithm has an additional negative penalty on the reward formulation if the grid constraint is violated as show in Equation (14) which is not present in the trading algorithm. The retrieved optimal winter battery schedule is shown in Figure 8a for the trading case and Figure 9a for the non-trading case. In comparison to the summer PV profile and tariff, it can be seen that rewards converge to very low values for the winter case. This is mainly attributed to the low PV profile and high winter grid tariffs for any energy purchased from the utility.



(**a**) (**b**)

Figure 7. Training curve showing reward convergences for episode number 0 to 20,000; (**a**) trading algorithm (**b**) non-trading algorithm.

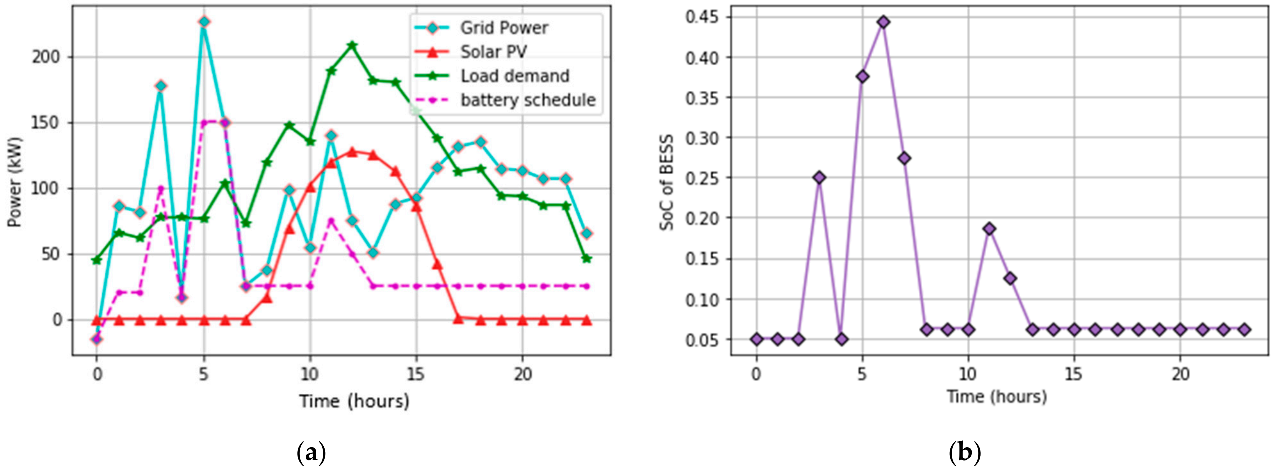


Figure 8. Observations without grid constraint; (**a**) Optimized energy schedule. (**b**) Tendency of the SoC of the BESS.

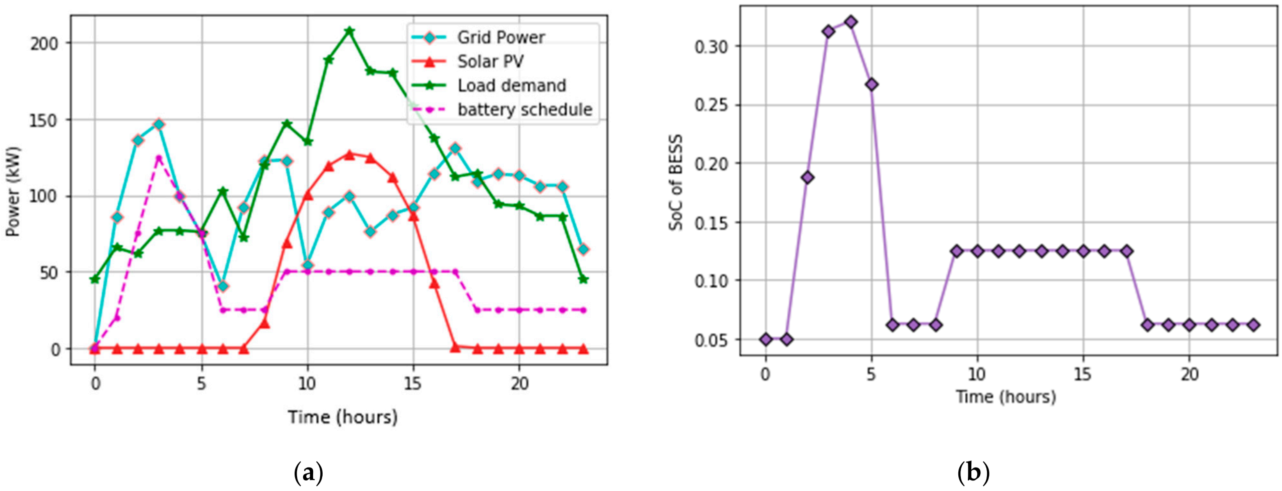


Figure 9. Observations with grid constraint; (**a**) Optimized energy schedule. (**b**) Tendency of the SoC of the BESS.

6.2.2. Energy Management

6.2.2.1. Results for Case Study 1 (Trading Algorithm)

Figure 8a displays energy schedules of the grid power and the battery plotted besides the winter solar PV output and grid tariff. As can be seen in Figure 2b the PV system produces small amount of energy between 08:00 and 16:00. Consequently, the trading operation will be limited as the total available PV generation will partly cover the micro-grid's demand. Also, the winter peak prices occur at 06:00 to 08:00 unlike the summer-time case where utility peak load starts at 07:00 [48]. In the trading algorithm, the EMS begins by scheduling zero grid power as the battery initial energy can fully meet the net load and later on a gradual increase in grid power utilization is seen. Between 00:00 and 05:00 the tariff is at its lowest, thus for about four hours very high-power absorption from the utility is recorded. When grid tariff increases during 06:00 to 09:00, the algorithm is seen to drastically lower the amount of power purchased from the main grid. At 11:00 when prices have reduced to mid-peak, the algorithm decides to increase power intake from the distribution network. From 13:00 to 23:00 the agent takes control actions of constantly charging the battery with 25 kW and supplying the remaining net load with power for the utility. In Figure 8b, it can be observed that the algorithm gradually increases the SoC of the battery up to 0.45 in the morning hours (04:00 to 06:00) to meet its net load as it anticipates the utility peak tariff which occurs from 06:00. As a result of raising the SoC, the algorithm is able to shift a large percentage of its net load until grid prices are reduced.

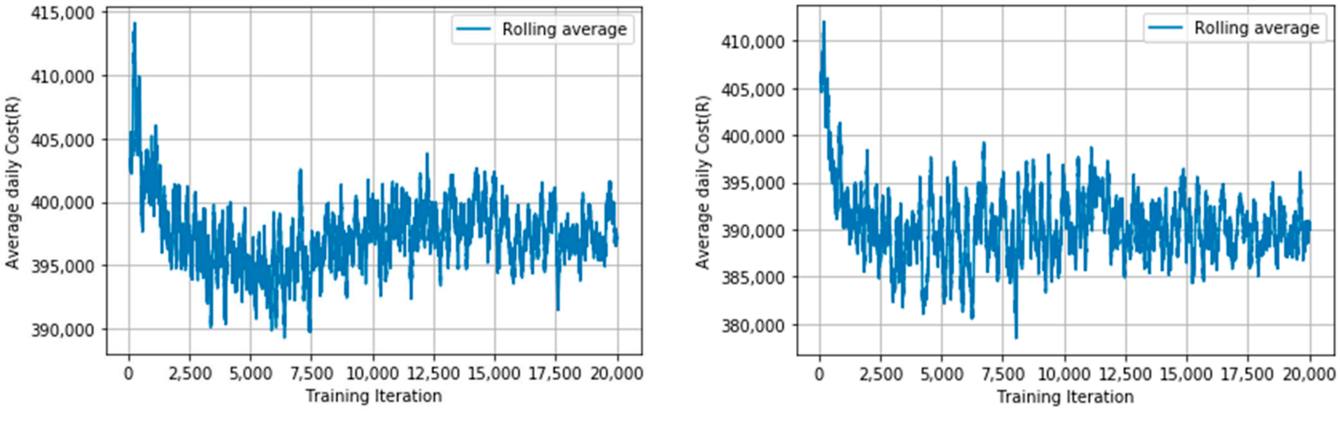
Unlike, during summertime it can be seen that battery utilization is rather low. As the PV is insufficient throughout the optimization horizon, the high deficit load computed by Equation (8) must be supplied by the utility grid. Thus from 13:00 the algorithm opts to keep the charge power as low as possible so as not to incur high cost of importing utility power to cover its deficit load and charge the battery. Similarly, the fact that the winter tariff is more expensive makes the algorithm to schedule lower charge energy so that the amount of power drawn from the grid is minimized. Finally, it can be observed that the SoC is maintained at its lowest level and the battery constraints are not violated.

6.2.2.2. Results for Case Study 2 (Non-Trading Algorithm)

Figure 9a presents the results of case study 2 which, as mentioned in Section 6.2.2.2, ensures that no energy is sold back to the utility grid. Between 00:00 and 05:00 the grid tariff is very low and PV power output is zero. a gradual increase in grid power utilization is observed as the algorithm chooses to charge up the battery with the cheap grid power so as to supply its net load as it would be optimal to do so. Between 06:00 to 09:00 the grid tariff shoots to its peak (R 278 increase in grid tariff is noticed in Figure 2b), the algorithm drastically lowers power intake from the utility and a constant charge power of 25 kW is recorded. At 09:00 to 17:00 when the tariff changes to mid-peak it can be observed that the algorithm slightly increases the battery charging power to 50 kW. A constant power intake by the battery is seen until the next peak tariff which occurs at 17:00 where the algorithm reduces the charging power to 25 kW. Observing Figure 8a, the trading algorithm learns to raise the SoC value to 0.45 unlike the non-trading which only reaches about 0.3 at 04:00, this causes the latter algorithm to only lower the grid power for one hour and later on rely heavily on the utility as the energy stored in the battery cannot support the microgrid's net load. However, in Figure 8a the power drawn from the utility is lowered for two consecutive hours during peak prices. In both cases, solar PV is very low and peak grid prices are also very high, however the algorithm learns to lower cost in these extremities.

6.2.3. Operational Cost during Winter

Figure 10a,b represent the average running cost variations during training. These curves tend to have almost similar characteristics although the non-trading plot is more erratic compared to the trading one. At the beginning of the training, the learning agent explores the action space and learns to avoid actions that result in high cost. In the final episodes actions that minimize cost are exploited for both cases. A final average global cost of about R 395,000 is recorded for both scenarios.



(**a**) (**b**)

Figure 10. Daily operational cost against the number of episode; (**a**) trading algorithm (**b**) non-trading algorithm.

6.3. Comparative Cost Study for Case 1 and Case 2

This section presents the energy cost comparison assessment for the optimized energy schedules. The comparison is based on the two case studies investigated, i.e., trading and non-trading cases, using both summer and winter PV and grid tariff data. Table 2 below shows the retrieved schedule energy cost for the two case studies in different season profile. In this problem, energy cost is the product of the power imported from the grid to cover microgrid's deficit power or/and charge the battery and the grid tariff. In the case of trading algorithm, the cost of exported energy is deducted.

$$Energy \quad cost = \sum_{t\,=\,1}^{t\,=\,24} G_t(t) P_{g\_p}(t) - \vartheta \, G_t(t) P_{g\_s}(t) \tag{25}$$

where $P_{g\_p}(t)$ denotes the power imported from the main grid, $P_{g\_s}(t)$ is the power exported to the utility grid, $G_t(t)$ is the instantaneous grid tariff and ϑ represents the selling price discounting factor. From the table it's apparent that in summer, the total operating costs are the lowest in both cases compared to the winter season. It can be reported that increasing PV generation would result to a much more profitable EMS operation in both the summer and winter seasons.

Table 2. Overview of energy cost for the optimal episode in the case studies considered.

Energy Cost	Summer Data (PV & Grid Tariff)	Winter Data (PV & Grid Tariff)
Trading Algorithm	R 103,708.71	R 367,322.73
Non-Trading Algorithm	R 107,891.05	R 375,403.00

To calculate the increase in percentage of total operating costs between the trading and non-trading case studies, Equation (26) is used.

$$I_{TC} \;=\; \frac{TC_{non-trading} - TC_{trading}}{TC_{trading}} \times 100 \tag{26}$$

I_{TC} denotes the increase of the total operational cost, (in percentage) $TC_{trading}$ and $TC_{non-trading}$ are the total operational cost of the trading and non-trading studied algorithms, respectively. The implementation of the proposed EMS for commercial load profile considering the no grid constraints (excess energy can be sold back to the utility) the total operating costs can reduce by 4.033% for summer data and 2.199% for winter data when compared to the non-trading algorithm. This phenomenon happens because, with the trading case there is more flexibility to feed power to the utility and earn some revenue whereas for the non-trading algorithm, less flexibility is experience by the agent when learning the environment as grid constraints cannot be violated. However, taking into account grid constraints is also technically beneficial particularly from the perspective of the local utility grid operators as the non-trading EMS avoids feeding any power back to the utility and this could lead to both technical and economic benefit to the microgrid owner and utility system operator.

Figure 11a–d below display the dispatching cost for both case studies for the retrieved optimized schedule.

Two case studies have been considered trading and non-trading settings. The objective is to reduce the total daily operation cost under the uncertainty of PV power, load demand, and grid tariff in both summer and winter seasons. Using numerical simulations and proper hyperparameter tuning, we confirmed that the proposed energy management schemes can efficiently minimize system operational costs (battery wear cost and cost of power purchased from the grid) under widely used south African time of use (ToU) grid tariff, and achieves desirable control actions which maximize solar PV usage while minimizing strain on the local utility during peak hours. The proposed energy management algorithm is intended to be applied in a number of intelligent grid environments including residential microgrids and smart energy facilities under different tariff structures to optimally schedule for energy consumption by efficiently managing the total energy produced and trading the surplus energy into the utility grid to make some profits.

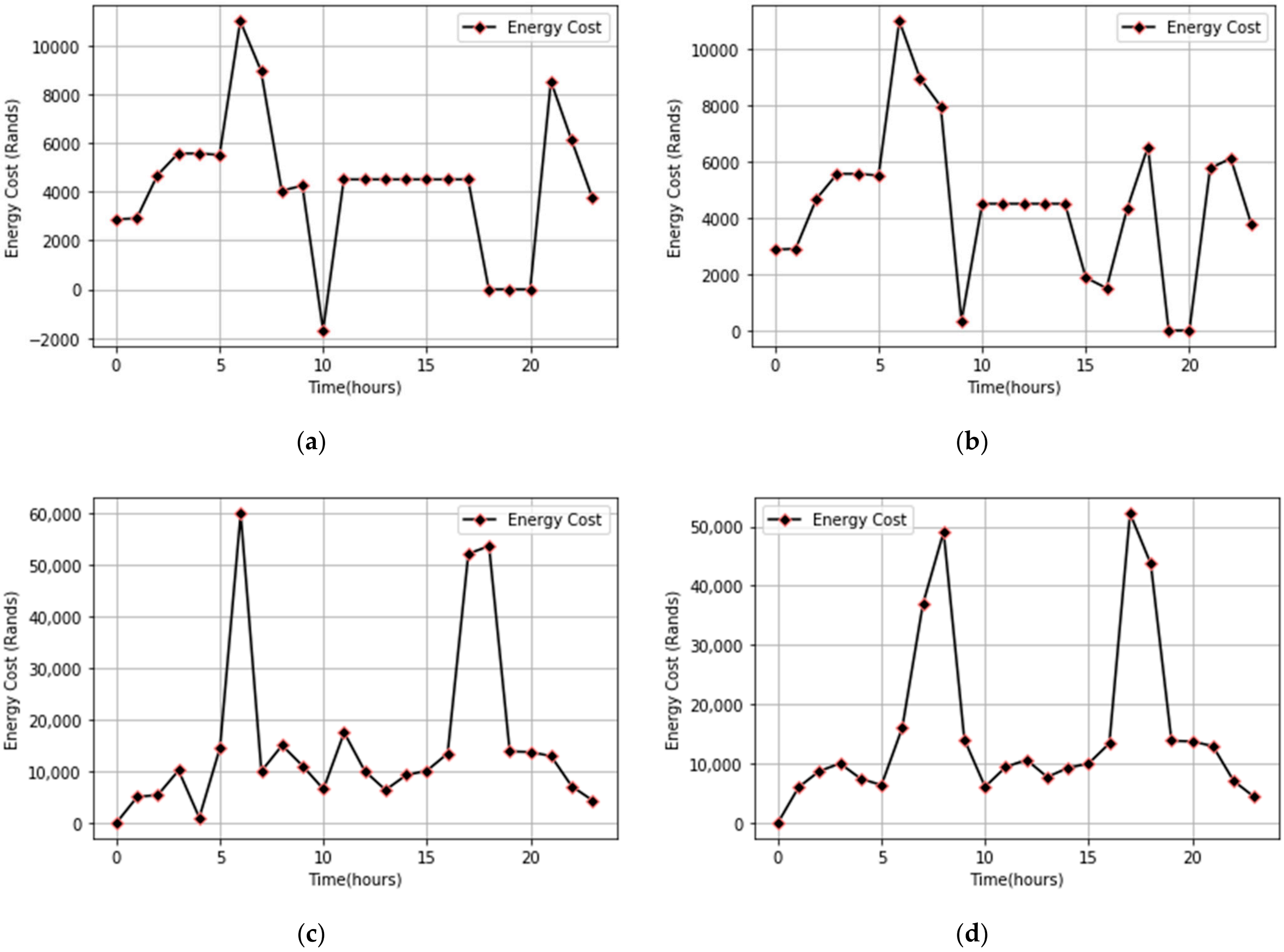


Figure 11. Cost profiles for the optimized schedule; (**a**) Summer dispatching cost profile for the optimized schedule (trading algorithm) (**b**) Summer dispatching cost profile for the optimized schedule (non-trading algorithm) (**c**) Winter dispatching cost profile for the optimized schedule (trading algorithm) (**d**) Winter dispatching cost profile for the optimized schedule (non-trading algorithm).

In comparison to the work presented in [50,51] the scheme proposed in this paper ensures that microgrid uncertainties to the utility (caused by stochastic nature of PV generation) are reduced where no excess PV is exported back to the utility. Still, the EMS is designed such that the battery can take advantage of the daily energy price fluctuations to buy the cheapest energy available during the period of low demand and sell it at the highest price. Also, a battery degradation model is embedded to ensure the battery is operated at power levels that do not significantly reduce its cycle life. Better battery utilization has also been achieved (for environments where peak PV production almost matches peak demand, and cases where PV is less compared to total load demand) thus supporting the grid shifting demand during peak load.

7. Conclusions

In this research, an energy management algorithm based on reinforcement learning was proposed for a grid-tied solar PV-battery microgrid supplying power to a commercial load. The novelty of the proposed work is mainly computational energy scheduling solutions for grid-tied microgrids in a highly stochastic setting where battery degradation model is also considered. To ensure practical application, the EMS is formulated as a Markov decision process considering state, action, and reward function. Considering a ToU grid tariff and stationary battery degradation cost, simulation results indicate that cost minimization has been achieved. Moreover, the findings show that responding appropriately to the dynamic grid tariff is a critical component of cost reduction and system efficiency. It is noteworthy that Q learning algorithm managed to lower operational costs in the two case studies regardless of the different tariff structures and the seasons considered. However, comparing the non-trading EMS to the trading EMS model, the energy trading algorithm achieved slightly better results as it reduced the energy costs by 4.033% more in summer season and 2.199% in winter season. The reinforcement learning approach successfully avoided high operational prices, efficiently utilized PV generation, and ensured reasonable SoC levels thus has the potential to be used in grid support applications such as peak load shaving and increasing system efficiency.

As future work, more scenarios can be explored using deep reinforcement learning techniques and other different grid tariffs. Also introducing flexibility on the demand side can be an interesting axis for future research. It is also to be noted that the behavior of some batteries can be characterized by rather complex relation between the SoC and the maximum power that the battery can deliver during discharge and, especially, absorb during charge. This issue can be explored further using elaborate techniques as proposed in this paper. This work has not considered thermal ageing of the battery and the reason for that is explained in Section 2.2. However, the proposed method can be further developed in future to include battery degradation due to thermal heating. The EMS is designed in a way that for different types of load curves, the algorithm can learn policies to increase the SoC of the BESS to meet the net demand independently or lower energy drawn from the grid at that time. Indeed, different load curves would affect the battery operation, but the designed algorithm is expected to handle the unexpected load peaks. However, a comparison of performance of the proposed method for slow, medium and fast demand fluctuations can be undertaken in future as a part of further uncertainty studies, which could not be covered in this paper due to unavailability of data for fast and slow demand fluctuations.

Author Contributions: Conceptualization, G.M.; Data curation, G.M.; Investigation, G.M.; Methodology, G.M.; Software, G.M.; Supervision, S.C.; Writing—original draft, G.M.; Writing—review & editing, S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would like to express their sincere appreciation to Mastercard Foundation and Electrical Engineering Department, University of Cape Town, Cape Town, South Africa for providing the sponsorship and infrastructure for this research.

Conflicts of Interest: The authors declare to the best of their knowledge that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Brown, H.E.; Suryanarayanan, S.; Heydt, G.T. Some characteristics of emerging distribution systems considering the smart grid initiative. *Electr. J.* **2010**, *23*, 64–75. [CrossRef]
2. Showers, S.O. Benefits and Challenges of Energy Storage Technologies in High Penetration Renewable Energy Power Systems. In Proceedings of the 2019 IEEE PES/IAS PowerAfrica, Abuja, Nigeria, 20–23 August 2019; pp. 209–214.
3. Luo, X.; Wang, J.; Dooner, M.; Clarke, J. Overview of current development in electrical energy storage technologies and the application potential in power system operation. *Appl. Energy* **2014**, *137*, 511–536. [CrossRef]
4. Badawy, M.O.; Sozer, Y.; Member, S. Power Flow Management of a Grid Tied PV-Battery System for Electric Vehicles Charging. *IEEE Trans. Ind. Appl.* **2017**, *53*, 1347–1357. [CrossRef]
5. Singhal, P.K. Dynamic Programming Approach for Large Scale Unit Commitment Problem. In Proceedings of the International Conference on Communication Systems and Network Technologies (CSNT), Katra, India, 3–5 June 2011; pp. 714–717.

6. Borra, V.S.; Debnath, K. Dynamic Programming for Solving Unit Commitment and Security Problems in Microgrid Systems. In Proceedings of the 2018 IEEE International Conference on Innovative Research and Development (ICIRD), Bangkok, Thailand, 11–12 May 2018; pp. 1–6.
7. An, L.N.; Quoc-Tuan, T. Optimal energy management for grid connected microgrid by using Dynamic programming method. In Proceedings of the 2015 IEEE Power & Energy Society General Meeting, Denver, CO, USA, 26–30 July 2015; pp. 1–5. [CrossRef]
8. Zhao, B.; Zhang, X.; Chen, J.; Wang, C.; Member, S.; Guo, L. Operation Optimization of Standalone Microgrids Considering Lifetime Characteristics of Battery Energy Storage System. *IEEE Trans. Sustain. Energy* **2013**, *4*, 934–943. [CrossRef]
9. Asefi, S.; Ali, M.; Gryazina, E. Optimal Energy Management for Off-Grid Hybrid System using Hybrid Optimization Technique. In Proceedings of the 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe), Bucharest, Romania, 29 September–2 October 2019; pp. 1–5. [CrossRef]
10. Hijjo, M.; Felgner, F.; Frey, G. PV-Battery-Diesel Microgrid Layout Design Based on Stochastic Optimization. In Proceedings of the 2017 6th International Conference on Clean Electrical Power (ICCEP), Santa Margherita Ligure, Italy, 27–29 June 2017; pp. 30–35. [CrossRef]
11. Jasmin, E.A.; Ahamed, T.P.I.; Jagathiraj, V.P. A Reinforcement Learning Algorithm to Economic Dispatch Considering Transmission Losses. In Proceedings of the TENCON 2008–2008 IEEE Region 10 Conference, Hyderabad, India, 19–21 November 2008; pp. 1–6.
12. Arwa, E.O.; Folly, K.A. Reinforcement Learning Techniques for Optimal Power Control in Grid-Connected Microgrids: A Comprehensive Review. *IEEE Access* **2020**, *8*, 208992–209007. [CrossRef]
13. Zia, M.F.; Elbouchikhi, E.; Benbouzid, M. Microgrids energy management systems: A critical review on methods, solutions, and prospects. *Appl. Energy* **2018**, *222*, 1033–1055. [CrossRef]
14. Sutton, R.S.; Barto, A.G.; Bach, F. *Reinforcement Learning: An Introduction*, 2nd ed.; The MIT Press: Cambridge, MA, USA, 2018; ISBN 9780262193986.
15. Mason, K.; Grijalva, S. A review of reinforcement learning for autonomous building energy management. *Comput. Electr. Eng.* **2019**, *78*, 300–312. [CrossRef]
16. Watkins, C.J.C.H.; Dayan, P. Technical Note: Q-Learning. *Mach. Learn.* **1992**, *292*, 279–292. [CrossRef]
17. Arwa, O.E.; Folly, K.A. Power Flow Management in Electric Vehicles Charging Station Using Reinforcement Learning. In Proceedings of the 2020 IEEE Congress on Evolutionary Computation (CEC), Glasgow, UK, 19–24 July 2020; pp. 1–8.
18. Mocanu, E.; Mocanu, D.C.; Nguyen, P.H.; Liotta, A.; Webber, M.E.; Gibescu, M.; Slootweg, J.G. On-line Building Energy Optimization using Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2019**, *10*, 3698–3708. [CrossRef]
19. Kim, S.; Lim, H. Reinforcement learning based energy management algorithm for smart energy buildings. *Energies* **2018**, *11*, 2010. [CrossRef]
20. Mbuwir, B.V.; Ruelens, F.; Spiessens, F.; Deconinck, G. Battery Energy Management in a Microgrid Using Batch Reinforcement Learning. *Energies* **2017**, *10*, 1846. [CrossRef]
21. Foruzan, E.; Soh, L.; Asgarpoor, S.; Member, S. Reinforcement Learning Approach for Optimal Distributed Energy Management in a Microgrid. *IEEE Trans. Power Syst.* **2018**, *33*, 5749–5758. [CrossRef]
22. Leo, R.; Milton, R.S.; Sibi, S. Reinforcement Learning for Optimal Energy Management of a Solar Microgrid. In Proceedings of the 2014 IEEE Global Humanitarian Technology Conference—South Asia Satellite (GHTC-SAS), Trivandrum, India, 26–27 September 2014; pp. 183–188.
23. Zeng, P.; Li, H.; He, H.; Li, S. Dynamic Energy Management of a Microgrid using Approximate Dynamic Programming and Deep. *IEEE Trans. Smart Grid* **2018**, *10*, 4435–4445. [CrossRef]
24. Arwa, O.E.; Folly, K.A. Energy Trading in Grid-connected PV-Battery Electric Vehicle Charging Station. In Proceedings of the Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA), Cape Town, South Africa, 29–31 January 2020; pp. 1–6.
25. Chen, P.; Liu, M.; Chen, C.; Shang, X. A battery management strategy in microgrid for personalized customer requirements. *Energy* **2019**, *189*, 116245. [CrossRef]
26. Chang, F.; Chen, T.; Su, W.; Alsafasfeh, Q. Control of battery charging based on reinforcement learning and long short-term memory networks. *Comput. Electr. Eng.* **2020**, *85*, 106670. [CrossRef]
27. Lu, R.; Hong, S.H. Incentive-based demand response for smart grid with reinforcement learning and deep neural network. *Appl. Energy* **2019**, *236*, 937–949. [CrossRef]
28. Lu, R.; Hong, S.H.; Zhang, X. A Dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach. *Appl. Energy* **2018**, *220*, 220–230. [CrossRef]
29. Nakabi, T.A.; Toivanen, P. Deep reinforcement learning for energy management in a microgrid with flexible demand. *Sustain. Energy Grids Netw.* **2021**, *25*, 100413. [CrossRef]
30. Kolodziejczyk, W.; Zoltowska, I.; Cichosz, P. Control Engineering Practice Real-time energy purchase optimization for a storage-integrated photovoltaic system by deep reinforcement learning. *Control Eng. Pract.* **2021**, *106*, 104598. [CrossRef]
31. Samadi, E.; Badri, A.; Ebrahimpour, R. Electrical Power and Energy Systems Decentralized multi-agent based energy management of microgrid using reinforcement learning. *Electr. Power Energy Syst.* **2020**, *122*, 106211. [CrossRef]
32. Shang, Y.; Wu, W.; Guo, J.; Ma, Z.; Sheng, W.; Lv, Z.; Fu, C. Stochastic dispatch of energy storage in microgrids: An augmented reinforcement learning approach. *Appl. Energy* **2020**, *261*, 114423. [CrossRef]

33. Ji, Y.; Wang, J.; Xu, J.; Fang, X.; Zhang, H. Real-time energy management of a microgrid using deep reinforcement learning. *Energies* **2019**, *12*, 2291. [CrossRef]
34. François-Lavet, V.; Taralla, D.; Ernst, D.; Fonteneau, R. Deep Reinforcement Learning Solutions for Energy Microgrids Management. In Proceedings of the 13th European Workshop on Reinforcement Learning (EWRL 2016), Barcelona, Spain, 3–4 December 2016; Volume 3–4, pp. 1–7.
35. Kuznetsova, E.; Li, Y.; Ruiz, C.; Zio, E.; Ault, G.; Bell, K. Reinforcement learning for microgrid energy management. *Energy* **2013**, *59*, 133–146. [CrossRef]
36. Cao, J.; Harrold, D.; Fan, Z.; Morstyn, T.; Healey, D.; Li, K. Deep Reinforcement Learning-Based Energy Storage Arbitrage with Accurate Lithium-Ion Battery Degradation Model. *IEEE Trans. Smart Grid* **2020**, *11*, 4513–4521. [CrossRef]
37. Ju, C.; Member, S.; Wang, P.; Goel, L.; Xu, Y. A Two-layer Energy Management System for. *IEEE Trans. Smart Grid* **2017**, *9*, 6047–6057. [CrossRef]
38. Abronzini, U.; Attaianese, C.; D'Arpino, M.; Di Monaco, M.; Genovese, A.; Pede, G.; Tomasso, G. Optimal energy control for smart charging infrastructures with ESS and REG. In Proceedings of the 2016 International Conference on Electrical Systems for Aircraft, Railway, Ship Propulsion and Road Vehicles & International Transportation Electrification Conference (ESARS-ITEC), Toulouse, France, 2–4 November 2016; pp. 1–6. [CrossRef]
39. Zhou, C.; Qian, K.; Allan, M.; Zhou, W. Modeling of the cost of EV battery wear due to V2G application in power systems. *IEEE Trans. Energy Convers.* **2011**, *26*, 1041–1050. [CrossRef]
40. Zhang, Y.; Zhang, T.; Wang, R.; Liu, Y.; Guo, B. ScienceDirect Optimal operation of a smart residential microgrid based on model predictive control by considering uncertainties and storage impacts. *Sol. Energy* **2015**, *122*, 1052–1065. [CrossRef]
41. Morales, E.F.; Zaragoza, J.H. An introduction to reinforcement learning. In *Decision Theory Models for Applications in Artificial Intelligence: Concepts and Solutions*; Sucar, L.E., Morales, E.F., Hoey, J., Eds.; IGI Global: Hershey, PA, USA, 2012; pp. 63–80. [CrossRef]
42. Huang, X.; Hong, S.H.O.; Member, S. Demand Response Management for Industrial Facilities: A Deep Reinforcement Learning Approach. *IEEE Access* **2019**, *7*, 82194–82205. [CrossRef]
43. Dewey, D. Reinforcement Learning and the Reward Engineering Principle. In Proceedings of the AAAI 2014 Spring Symposium, Palo Alto, CA, USA, 24–26 March 2014; pp. 1–8.
44. Shin, M.; Choi, D.; Kim, J. Cooperative Management for PV/ESS-Enabled Electric Vehicle Charging Stations: A Multiagent Deep Reinforcement Learning Approach. *IEEE Trans. Ind. Inform.* **2020**, *16*, 3493–3503. [CrossRef]
45. Kearns, M. Near-Optimal Reinforcement Learning in Polynomial Time. In *Machine Learning*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2002; Volume 49, pp. 209–232.
46. Alwan, H.O.; Sadeghian, H.; Abdelwahed, S. Energy Management Optimization and Voltage Evaluation for Residential and Commercial Areas. *Energies* **2019**, *12*, 1811. [CrossRef]
47. Global Solar Atlas. Available online: https://globalsolaratlas.info/detail?c=-33.928992,18.417396,11&r=ESP&s=-33.928992,18.417396&m=site&pv=medium,0,29,300 (accessed on 8 February 2021).
48. Tariff History. Available online: https://www.eskom.co.za/CustomerCare/TariffsAndCharges/Pages/Tariff_History.aspx (accessed on 1 March 2021).
49. Lithium-Ion Battery Pack Costs Worldwide between 2011 and 2030. Available online: https://www.statista.com/statistics/883118/global-lithium-ion-battery-pack-costs/ (accessed on 3 February 2021).
50. Bracco, S.; Brignone, M.; Delfino, F.; Girdinio, P.; Laiolo, P.; Procopio, R.; Rossi, M. A simple strategy to optimally design and manage a photovoltaic plant integrated with a storage system for different applications. In Proceedings of the AEIT International Annual Conference, Cagliari, Italy, 20–22 September 2017; pp. 1–6.
51. Delfino, F.; Ferro, G. Sustainable Energy, Grids and Networks Identification and optimal control of an electrical storage system for microgrids with renewables. *Sustain. Energy Grids Netw.* **2019**, *17*, 100183. [CrossRef]