Efficient use of PV in a Microgrid using Reinforcement Learning

Khawaja Haider Ali[1,*], Asif Khan[2],

[1] Department of Electrical Engineering, Sukkur IBA University, Sukkur, Pakistan
[2] Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan
[*]Corresponding author: haiderali@iba-suk.edu.pk

Abstract

Artificial Intelligence is a new concept to optimize or schedule the energy storage system of the Microgrid. The reinforcement learning (RL) method can be used in the effective scheduling of the battery connected to the microgrid. The proposed strategy aims to reduce energy costs while prioritizing both energy balance and user comfort within the microgrid. The key innovation lies in developing an optimal policy for battery actions (charging, discharging, idle) using a model-free stochastic approach. One significant aspect that sets this work apart from others is its acknowledgment of the non-deterministic nature of the state of charge (SOC) of the battery. Unlike systems that solely rely on grid charging, our approach takes into account the unpredictability of renewable energy sources, particularly solar power, which heavily depends on varying time instances and weather conditions throughout the day. Consequently, the SOC of the battery exhibits non-deterministic behavior due to the uncertainty in the availability of excess renewable energy for charging. The RL-based policy presented in this research capitalizes on the effective utilization of photovoltaic sources, optimizing the battery's discharge and idle states. By intelligently adapting to the dynamic energy supply from renewable sources, the proposed approach ensures that the battery is charged only when surplus energy is available beyond fulfilling the overall system load demand.

Keywords—AI, Reinforcement Learning, Scheduling, Optimization, Charging, Discharging, Battery

1 Introduction

There are different ways to optimize the microgrid to get the maximum benefit out of it [1]. Management of microgrids will also lead to cutting down the cost which can directly affect positively to the distributor and customers [2]. Working on a microgrid in a more efficient way increases the overall profit of the generation companies [3]. There are two modes in which microgrid can operate:

1) Off Grid Mode
2) On Grid Mode

In both above cases, if it is managed optimally, it will pay a lot in the long run for both providers and users. In the last few years, a lot of research has been done to control, operate, and distribute energy via microgrid Optimally [4-7]. There are different algorithms and strategies suggested to manage the whole microgrid. For example; selection of renewable energy source, forecasting of the load and demand, sizing of the storage system, Scheduling of storage devices, sitting, and many more. These plans mainly consist of model-based and non-model-based approaches. It is not necessary that optimization problems can be handled by one type of approach every time. It is a fact that in different kinds of Scenarios, optimization can be done by different methods. This is because of the nature of the problem, some time it is due to the stochastic behavior of the system or due to deterministic and non-deterministic characteristics of the environment. The algorithms used in optimization techniques are linear programming, MPDP, Game theory, and so on [8].

In this work, the model-free Markov's decision process technique will be applied to schedule the storage device of the microgrid. The microgrid is connected to the main grid. Load, photovoltaic (PV) and tariff profiles are important to know in 1st step to optimize and manage the microgrid. It is very important to know the demand timing of the load. For example, peak, off-peak, and medium peak. Load can be varied at different times of the day. If renewable sources are present within a system or microgrid, its output is

ISSN: 2523-0379 (Online), ISSN: 1605-8607 (Print)
DOI: https://doi.org/10.52584/QRJ.2101.13

dependent on the different conditions (e.g., weather) [8]. Therefore, microgrids supply power either from a renewable source or from the main grid to load. The presence of the battery or energy storage system in a microgrid has many advantages. One of the advantages of a battery can be; supplying power to load in case of absence or shortage of power from renewable sources or generating units. Moreover, if the microgrid did not have so much demand at the load side then the energy of the battery could be injected into the main grid as well. But, to do this battery or microgrid should know the state of charge of the battery.

For the sake of management, it is also vital for the battery to charge itself from the utility grid when the tariff is low or from a renewable source connected to the microgrid. In this way charged batteries can be utilized optimally during the need. (when the tariff is high, or load demand is higher than renewable production). To do so, the state of charge (charging and discharging) of the battery should be controlled according to its optimum level [9]. The battery should be self-efficient to know when to charge and where to charge either from the renewable source or from the main grid (when the tariff is low). It is also part of the management system of the battery to discharge up to the optimum level when there is demand from the load side. In other cases, it can deliver or sell its power to the main grid.

Battery should take care while supplying its energy to main-grid that after discharging, it should get lo tariff from the main grid or get power from a renewable source to charge again. Otherwise, do not sell energy to the main grid to save its charging for the next day or high-demand load conditions. This can be done if the battery manages itself on its own by adopting any algorithm such as machine learning. It is suggested in this work that the battery will learn itself by knowing load demand, tariff rates, weather conditions, and other necessary variables to become self-efficient and make decisions on its own. For example, when to charge where to charge, when to discharge, and where to dispatch its energy. The battery is charging only from a Renewable source, which is photovoltaic (PV) in this case with the condition of remaining available energy from it after fulfilling the demand of the load. Then Reinforcement learning is used to determine the best possible action of the battery, which is either discharging or remaining Idle.

1.1 Problem Statement of this Work

Find the best optimal action charging, discharging, and idle mode of the battery by analyzing the availability of PV, storage energy (SE) of the battery, and cost of the utility grid.

1.2 Objectives

The aim of this work is to optimize the microgrid in such a way that it becomes cost-efficient, with a primary focus on reducing energy bills based on the following parameters:

1) Comfortable and reliable system for the users.
2) Maximum utilization of Renewable energy resources.
3) Less dependent on the Main grid.
4) Increase the cost savings.
5) Proper charging and discharging of the battery by considering the important parameters of the battery.

2 MICROGRID MODEL

Figure 1 depicts the proposed model in an ”on-grid” mode, also known as the connected mode. This architecture offers the potential to significantly enhance system autonomy while reducing dependencies on the main grid power. By operating in this mode, the locally generated energy is consumed directly within the microgrid, without the need to inject excess energy back into the grid [9]. This approach facilitates a more self-sustaining and efficient d management system, leading to reduced reliance on external grid resources and promoting a higher degree of energy independence for the microgrid.

This model helps to make great use of renewable energy. As, it integrates a storage energy system to store the energy, during the day and to use it during the peak to fulfill the requirement of the load. Figure 1 is given below composed of renewable energy sources (PV), the storage energy system, the main grid energy, and the load. Cost is one of the important factors, that has been considered as a priority at every perspective. As the tracking system is much cost cost-effective comparatively fixed mount panels, it is wise to use trackers to get maximum output on the lesser amount. The below results depict the performance of Solar panels on the basis of different parameters and technologies.

2.1 Proposed Method

Reinforcement Learning (RL) is an advanced machine learning technique utilized in this algorithm to make decisions regarding the operational mode of a battery system. The primary objectives are to minimize the

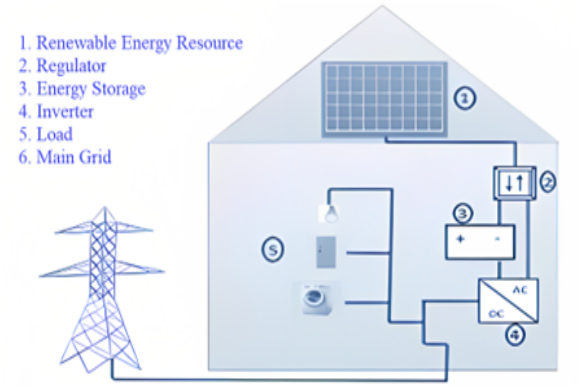

Fig. 1: Illustration of the Microgrid System Mode

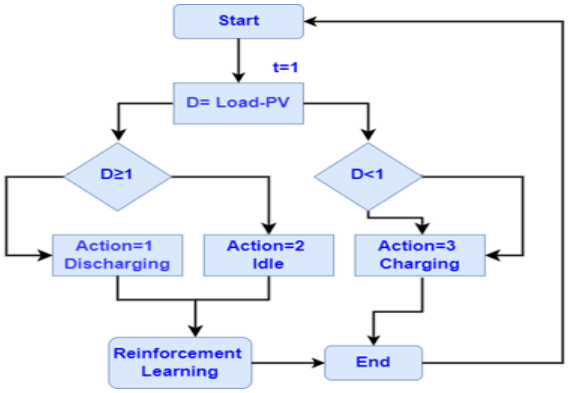

Fig. 2: Comprehensive Flow Chart of the Proposed Model

cost of electricity and maximize the consumption of locally produced electricity. To achieve this, the RL algorithm employs a sequential decision-making approach based on Markov Decision Processes (MDP). The fundamental idea behind RL is to find the optimal actions for the battery system by learning from interactions with the environment. This learning process is carried out through a policy represented by the Q-function. The goal is to maximize the Q-value function, which estimates the expected future reward for a given state-action pair. To facilitate this learning, the Q-function is typically organized in a tabular format known as the Q-table. This table contains entries for each state-action pair and represents the potential future rewards. During the RL process, the agent explores the environment by taking actions randomly to gather information about the rewards associated with different states and actions. This exploration phase helps update the Q-table by gradually refining the estimates of the Q-values based on the observed rewards [10]. As the agent continues to interact with the environment and update the Q-table, it gradually acquires knowledge about the best actions to take in each state.

Once the Q-table has been sufficiently learned and updated, the agent transitions from exploration to exploitation. At this stage, the agent leverages the information stored in the Q-table to select the best actions for each state, which are expected to lead to higher rewards and lower electricity costs. The process of updating the Q-table is typically done using the Q-Learning algorithm, which iteratively updates the Q-values based on the Bellman equation. This equation combines the current reward received from an action and the estimated future rewards from the next state, weighted by a discount factor. This iterative update process refines the Q-values and leads the agent towards making better decisions over time [11]. Initially, we explore (process of exploration) the environment and update the Q-Table. When the Q-Table is ready, the agent will start to exploit the environment and start taking better actions [11][12].

$$(s,a) = Q(s,a) + \alpha(Reward + Gamma * max(Q(s^{'},a)) - Q(s,a)) \quad (1)$$

Figure 2, given below illustrates the comprehensive methodology employed in this study, presenting various facets that have been explored. The diagram showcases the diverse dimensions of our research and the approach taken to investigate the subject matter thoroughly.

3 FEATURES OF THE IMPLEMENTED ALGORITHM (RL)

3.1 SYSTEM STATES

In RL, it is a very crucial part to define the states of the system. As states decide the maximum horizon of the model. Also, careful declaration of states makes the model close to practical. Two types of problems can be dealt with in RL methodology, finite horizon or infinite horizon. This work gives a finite horizon problem solution. We are interested in getting the optimal solution to our problem in one complete day. The day is divided into 24 hours and further, it is divided among hour time intervals. So, there are 24 states at a time. The generalized equation of system states is:

$$S = S_t * SSE \quad (2)$$

where S_t is the time feature of the state which is divided into 24 intervals. As, one day have 24 hours. SSE, represents the states of the storage energy in the

battery. In this work, three SSE states are considered such as:

$$S_{SE}min \leq S_{SE} < S_{SE}min + \frac{1}{3}capacity \quad (3)$$

$$S_{SE}min + \frac{1}{3}capacity \leq S_{SE} < S_{SE}min + \frac{2}{3}capacity \quad (4)$$

$$S_{SE}min + \frac{2}{3}capacity \leq S_{SE} < S_{SE}max \quad (5)$$

where

$$capacity = S_{SE}max - S_{SE}min \quad (6)$$

At, every time interval t there can be possibility of one SSE level out of three S_{SE}, $S_{SE}2$ or $S_{SE}3$. Therefore,

$$t(K) = S_{SE}1, S_{SE}2, S_{SE}3; K = 1, 2, ..., 24$$

Hence, Equation 2 gives, total number of states used in this work:

$$S = S_t \cdot S_{SE} = 24 \cdot 3 = 72$$

3.2 Actions

Depending on the actual state out of 72 states, the system chooses between the following actions.

$$A = [0, 1, 2] \quad (7)$$

$a0$ = Charging of battery when PV is there after fulfilling the demand of load; $a1$ = discharge battery; $a2$ = Battery is in idle mode

3.2.1 Scenario's for a1 (discharging)

1) Complete the load demand by using battery only if SE of the battery is enough.
2) If SE of battery is not enough then battery and PV used to fulfil the demand
3) If SE of battery is not enough then battery and Main grid is used to complete the demand of load

All above conditions are evaluated by the cost factor as lesser cost give higher reward. So higher reward tells the agent which is the most suitable case is $(C1, C2, andC3)$ to discharge the battery.

3.2.2 Scenario's for a2 (idle)

Load demand is fulfilled by PV or (PV+ grid) or grid alone according to priority. Therefore, in our work the-most suitable case is to fulfil the demand of load from PV. If PV is not enough then (PV+ main grid) will be the option. Otherwise, main grid will fulfil the demand of load.

3.3 States VS Actions

In each state out of 72 we have 1 action which means that at every state depending upon the SE of the battery actions may different. As, at every time step t, SE of the battery is checked and depending upon SE level (state) the actions are noted. Here are three states of SE so at every time interval t so, there are three sub states of SE as mentioned in equations 3, 4 and 5 in which actions are taken. The actions at time 1 may be different in each round (steps/iteration) because of the SSE different levels. This will be updated in every round depending upon the reward function. As the system will learn with continuous iteration and in last iteration system converges to give action per state/time which is the best or optimal. At the end when system converges, we extract the optimal actions (24) out of 72 in such a way that initial SE of the battery is set to see the upcoming actions of the battery within 24 hour's time interval.

3.4 Reward and cost Function

The reward is inversely proportional to cost function ($Reward = 1/cost$). So, higher cost makes the reward lesser and vice versa. Cost function have following below mentioned variables expressed in equation 8 and 9.

$$Cost(s, a) = P(t)\cdot(D)t) - min(b, S_{SE}(t) - S_{SE}min)); \; if \; a = 1 \quad (8)$$

$$Cost(s, a) = P(t) \cdot (D)t) \; if \; a = 2 \quad (9)$$

where, $P(t)$ represents the Price of energy from the main grid. $D(t)$ is the difference between load and PV power (net demand). $b = S_{SE}max/10$ which express the maximum $charging/discharging$ rate. $S_{SE}min = 0.3 * S_{SE}max$. The initial SSE of the battery is equal to $S_{SE}max$ which is 5.28 here in our model.

3.5 Exploration VS Exploitation

In the RL setting, batch of data is not provided like in supervised learning. In case of RL as we go along during training, we're gathering data. While the actions we take can affect the data. Therefore, sometimes it makes sense to take different actions to observe or gather new data [11]. The epsilon(ϵ) considered in this paper is as below:

$$\epsilon = \frac{\epsilon}{\sqrt{((M - M_{max}))}} \quad (10)$$

Above function allows the algorithm to dynamically adjust the learning rate, ensuring that the network effectively learns and optimizes its performance while avoiding the pitfalls of both overly slow and excessively fast learning rates.

3.6 Learning rate

The learning rate, often referred to as the alpha value, plays a crucial role in determining how rapidly a neural network replaces old beliefs with new ones [13]. Striking the right balance is essential, as the objective is to discover a learning rate that is sufficiently low to converge towards valuable information while avoiding excessively lengthy training periods [14]. In this study, the α value is represented by the following equation:

$$\alpha = \frac{\alpha}{(N - N_{max}} \tag{11}$$

3.7 Discount factor

The discount factor refers to a numerical value that represents the contrast between rewards received in the future and rewards obtained in the present [15]. In other words, it is a parameter used in various decision- making processes and mathematical models. When considering rewards over time, the discount factor helps account for the concept of time preference, where individuals tend to value immediate rewards more than delayed ones. It allows for the comparison and evaluation of rewards occurring at different points in time, helping to make rational decisions when faced with choices involving potential gains or losses at various time intervals.

3.8 Algorithm of Reinforcement Learning

Figure 3 illustrates the step-by-step process of reinforcement learning algorithms. The agent then observes the current state of the environment and selects an action based on its chosen policy. The chosen action is executed in the environment, resulting in a new state and a corresponding reward. After receiving the reward, the agent updates its policy or value functions to improve decision-making in future interactions. This iterative process continues until the agent's learning converges or reaches a predefined stopping point.

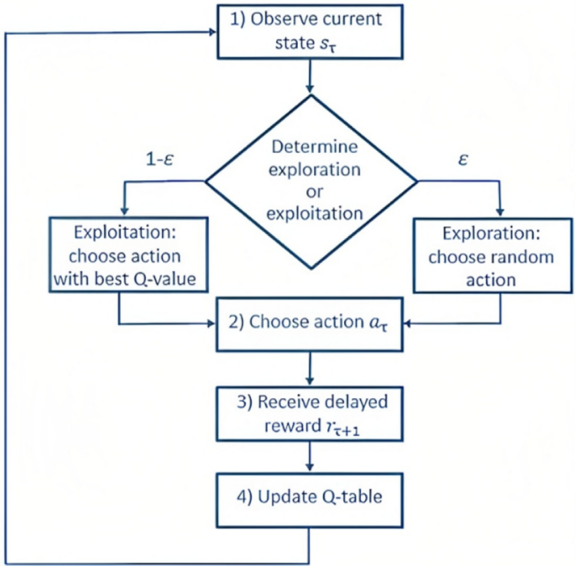

Fig. 3: The Q-learning algorithm, represents the fundamental framework used to optimize decision-making in this study

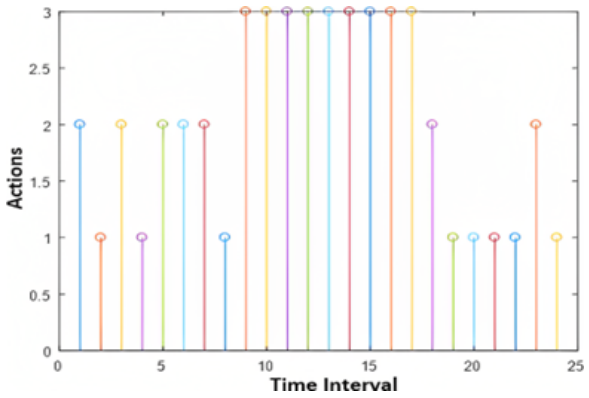

Fig. 4: Optimal actions of battery in 24 hours' time horizon

4 RESULT AND DISCUSSIONS

The graphs and algorithm of RL is made using Matlab version R2019a. The outcomes of our proposed algorithm applied in above mentioned model will be discussed in detail below.

4.1 Optimal actions of battery w.r.t 24 hours' time

This study encompasses a total of 72 different states, which play a significant role in evaluating the optimal actions within a 24-hour time frame. The 24-hour period is subdivided into one-hour time steps, allowing for a detailed analysis of actions at each step. After the algorithm reaches its convergence point, the optimal actions are extracted, with the intention of executing them in real-time on a physical microgrid system. The main objective of this work is to provide an optimization solution for the microgrid, specifically aiming at cost savings over the course of a single day. For every hour within this day, the agent acquires the optimal action for controlling the battery.

In Figure 4 below, the optimal actions achieved after the convergence of the Reinforcement Learning (RL) algorithm is visually presented. The vertical axis (y-axis) represents the different actions available to the agent for battery management. Specifically, the numbers on the y-axis have the following interpretations:

1) The value 1 on the y axis indicates the action of

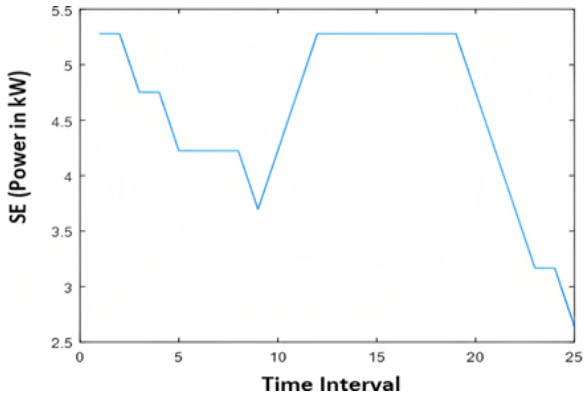

Fig. 5: SE level of the battery after Convergence

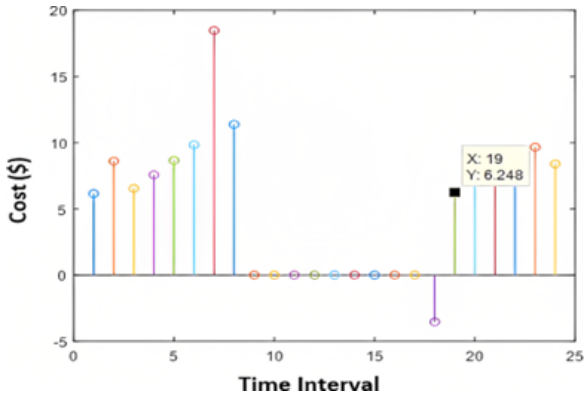

Fig. 6: Optimal cost after convergence of the algorithm

discharging the battery, which involves utilizing the stored energy to power the microgrid or meet electricity demands.

2) The value 2 represents the idle mode, where the battery remains inactive without any charging or dis-charging, conserving its current state of charge.
3) The value 3 illustrates the action of charging the battery, which involves replenishing the stored energy reserves for future use.

4.2 SE level of battery after convergence

Figure 5 depicts a graphical representation showcasing the variations in State of Charge (SOC) of a battery concerning different Self-Discharge Levels (SE) observed during the post-convergence phase of an algorithm. It refers the amount of electrical energy stored in the battery at a given point in time, while SE indicates the rate at which the battery discharges itself without any external load connected. The graph (shown in Fig. 5) illustrates how the State of Charge of the battery changes over time as the algorithm reaches its convergence point. It provides valuable insights into the behavior of the battery self-discharge at different stages of algorithmic convergence.

4.3 Optimal cost after convergence

Figure 6 displays the optimal cost achieved in the final iteration of the optimization algorithm. The x-axis represents the iteration number, showcasing the sequential steps the algorithm took during its optimization process. The y-axis represents the corresponding optimal cost values obtained by the algorithm at each iteration.

Figure 6 given above shows algorithm progresses through its iterations, it continuously refines its solution to minimize the cost metric associated with the microgrid's operation. The goal is to achieve the most cost-effective solution that addresses specific aspects of the microgrid, such as energy consumption, battery usage etc. The graph visually illustrates the trend of the cost values over the iterations, enabling us to assess the algorithm's convergence. A decreasing trend in cost values indicates that the algorithm is approaching its optimal solution, with successive iterations resulting in progressively lower costs. The point on the graph representing the last iteration's cost value is considered the "optimal cost," signifying the best solution found by the algorithm after reaching convergence.

4.4 Validity of the proposed optimal solution

The solution of optimizations depend on the different scenarios. Some time it is not possible to solve a problem from all techniques of optimization. One can check the validity of the solved problem by analyzing the different parameters of the applied algorithm. In this work the optimal policy of the actions got after convergence can be checked by applying some basic concepts of RL. To validate this work below are some justifications of this approach which shows that this AI technique solve the optimization problem related to microgrid system.

4.4.1 Sum of Reward behaviour VS every episode/round

In RL algorithm, it is most important fact that if reward has increasing trend with respect to each episode then it means the agent is learning continuously. Also, the purpose of epsilon greedy policy also suggests that in the beginning when epsilon value is close to 1 the agent has tendency to explore the environment [16]. That is why reward values fluctuate during learning or changes from one episode to another. But when epsilon becomes lower and lower until it reaches to value equal to zero, the agent start exploitation [17].

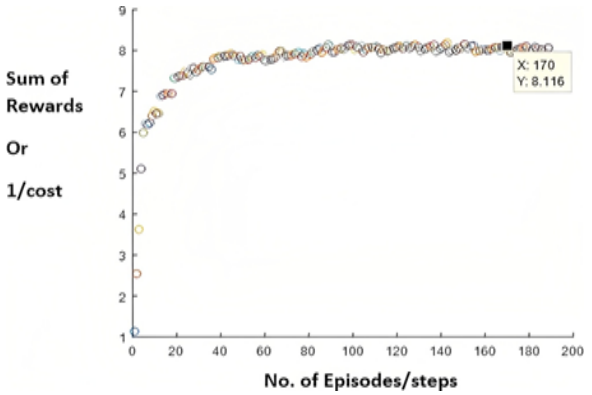

Fig. 7: A detailed graphical representation of the Sum of Rewards obtained after each episode during the experiment

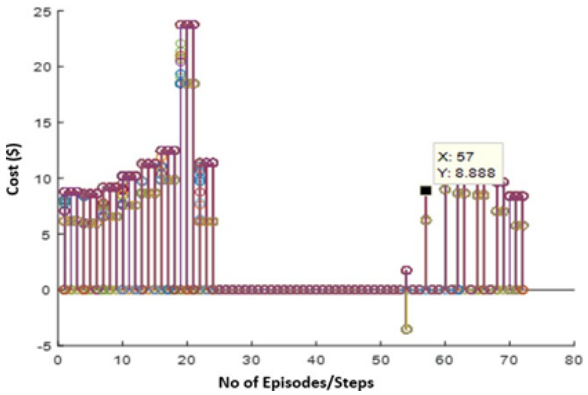

Fig. 8: Cost function after each episode

It means that the system had learned from its past experiences and form these experiences the best option becomes the optimal solution. The optimal solution has approximately maximum reward than previous iterations or episodes.

Therefore, sum of rewards in last episode should be greater than sum of rewards in individual previous episode. In this work the validity of the optimization had been checked by this approach as well. The Figure 7 below, shows that the sum of reward in last episode that are approximately greater than all previous episodes.

Figure 7 given above illustrates the learning progress and optimization validation of a reinforcement learning agent through multiple episodes of interaction with the environment. In each episode, the agent takes actions and receives rewards based on its decisions. The x-axis of the graph represents the episodes, showing the sequential order in which the agent's interactions occur. As we move from left to right along the x-axis, we traverse the timeline of the experiment. The y-axis displays the cumulative Sum of Rewards obtained by the agent at the end of each episode. This cumulative sum represents the total rewards accumulated by the agent during its interactions from the beginning of the experiment up to that particular episode.

4.4.2 Cost VS each episode

The graph in below figure 8 compares the cost of each action taken during many episodes/steps. The cost in last episode has lower value than previous ones. The target of this optimization was to reduce the cost as well. It means the power getting from the main grid becomes lower and lower as the agent learns more during the different episodes. Therefore, the system learns that the most efficient way to reduce cost is to become less dependent on utility grid.

The main objective of the optimization process is to minimize the overall cost incurred by the system. The Figure 8 shows a consistent decreasing trend in the cost values as the episodes progress. This trend indicates that the optimization process is successful in reducing the overall cost of system operations. Furthermore, the decreasing cost trend suggests that the system is becoming less dependent on power obtained from the utility grid. The agent's learning process allows it to optimize energy usage and explore alternatives, resulting in reduced reliance on potentially more expensive grid-supplied electricity. As the agent gains experience through different episodes, it becomes more efficient in selecting actions that lead to cost savings. The graph demonstrates the agent's improved learning efficiency, as evidenced by the declining cost values over time.

5 Conclusion and Future Work

AI had been used in many applications in past few years. RL is one of the promising algorithms in AI field. RL is mostly used in game theory [18]. The idea of using this approach in the management of microgrid is quite novel. This technique applied here to find best optimal policy for the actions of battery taken in each hour of the day. At the end of the result section, optimal actions of the battery are verified by applying theoretical concepts of RL which proves authenticity of this work. The scope of this work had many dimensions. As, it can be referred to apply and solve more complex energy management problems at real time near future. This is very unique idea to optimize the microgrid especially when energy storage system attached to it. This optimization technique has an ability to learn like humans [19]. So, learning process in RL can be enhanced or more optimized by increasing

the steps (No. of Episodes). Also, the learning procedure of this algorithm depends on Epsilon and Alpha selection [20][21]. So, by selecting these two parameters accurately, optimal actions after the convergence of an algorithm become more cost effective. Some more useful steps can be taken in future work to make it more practical.

- In future, the optimization of the problem discussed in this work can be enhanced by charging battery from the main grid as well. For example, if the tariff of the main grid is less, than battery can be charged and utilized in high tariff time.
- Parameter which is State of charge of the battery (SOC) is important in terms of battery life. It can be considered along with SE or without it in future.
- The inverter model can also be made in future to make this algorithm more practical. Also, SOC states can be defined by checking the original parameters of the battery used in this microgrid structure.
- This technique can be tried on real system on run time. The load, PV and tariff profiles are also used accurately according to the region where Microgrid is installed.
- In future, this algorithm can be validated by any other optimal technique such as by linear programing or by solving the problem by other MDP techniques. Comparison of both solutions give more accurate idea about the validity of RL in different optimization problems.

References

[1] Thirunavukkarasu, G. S., Seyedmahmoudian, M., Jamei, E., Horan, B., Mekhilef, S., & Stojcevski, A. (2022). Role of optimization techniques in microgrid energy management systems—A review. Energy Strategy Re-views, 43, 100899.

[2] Shahzad, S., Abbasi, M. A., Ali, H., Iqbal, M., Munir, R., & Kilic, H. (2023). Possibilities, Challenges, and Future Opportunities of Microgrids: A Review. Sustainability, 15(8), 6366.

[3] Chartier, S. L., Venkiteswaran, V. K., Rangarajan, S. S., Collins, E. R., & Senjyu, T. (2022). Microgrid emergence, integration, and influence on the future energy generation equilibrium—A Review. Electronics, 11(5), 791.

[4] Ishaq, S., Khan, I., Rahman, S., Hussain, T., Iqbal, A., & Elavarasan, R. M. (2022). A review on recentdevelopments in control and optimization of micro grids. Energy Reports, 8, 4085-4103.

[5] Roslan, M. F., Hannan, M. A., Ker, P. J., Mannan, M., Muttaqi, K. M., & Mahlia, T. I. (2022). Microgrid control methods toward achieving sustainable energy management: A bibliometric analysis for future directions. Journal of Cleaner Production, 348, 131340.

[6] Abd ul Muqeet, H., Munir, H. M., Ahmad, A., Sajjad, I. A., Jiang, G. J., & Chen, H. X. (2021). Optimal operation of the campus microgrid considering the resource uncertainty and demand re-sponse schemes. Mathematical Problems in Engineering, 2021, 1-18.

[7] Ahmad, S., Shafiullah, M., Ahmed, C. B., & Alowaifeer, M. (2023). A Review of Microgrid Energy Management and Control Strategies. IEEE Access.

[8] S.Bahramirad "Reliability-constained optimal sizing of energy storage system (ESS) in a microgrid", in IEEE transactions on smart grid, volume.3 no.4, December 2012.

[9] Essayeh, Chaimaa, Mohammed Raiss El-Fenni, and Hamza Dahmouni. "Cost-Effective Energy Usage in a Microgrid Using a Learning Algorithm." Wireless Communications and Mobile Computing 2018 (2018).

[10] Brida V. Mbuwir"Battery Energy Management in a Microgrid Using Batch Reinforcement Learning †", 15 October 2017; Accepted: 7 November 2017; Published: 12 November 2017.

[11] https://www.quora.com/What-is-the-learning-rate-in-neural-networks retrieved on dated 12/02/22.

[12] https://towardsdatascience.com/reinforcement-learning-demystified-exploration-vs-exploitation-in-multi- armed-bandit-setting-be950d2ee9f6 retreived on dated 09/06/22.

[13] Kofinas, Panagiotis, George Vouros, and Anastasios I. Dounis. "Energy management in so-lar microgrid via reinforcement learning using fuzzy reward." Advances in Building Energy Re-search 12.1 (2018): 97-115.

[14] Essayeh, Chaimaa, Mohammed Raiss El-Fenni, and Hamza Dahmouni. "Optimal Energy Exchange in Micro-Grid Networks: Cooperative Game Approach." 2018 Renewable Energies, Pow-er Systems & Green Inclusive Economy (REPS-GIE). IEEE, 2018.

[15] Avijit Das ; Zhen Ni, A Computationally Efficient Optimization Approach for Battery Systems in Islanded Microgrid, IEEE Transactions on Smart Grid (Volume: 9 , Issue: 6 , Nov. 2018).

[16] Ali, Khawaja Haider, Mohammad Abusara, Asif Ali Tahir, and Saptarshi Das. 2023. "Du-al-Layer Q-Learning Strategy for Energy Management of Battery Storage in Grid-Connected Mi-crogrids" Energies 16, no. 3: 1334. https://doi.org/10.3390/en16031334

[17] Tu A. Nguyen "Stochastic Optimization of Renewable-Based Microgrid Operation Incor-porating Battery Operating Cost", IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 31, NO. 3, MAY 2016.

[18] Kuznetsova, Elizaveta, et al. "Reinforcement learning for microgrid energy management." Energy 59 (2013): 133-146. Optimal Sizing of Smart Grid Storage Management System in a Mi-grogrid. Bahramirad S., Daneshi H. s.l. : IEEE, 2011. IEEE. pp. 1-6.

[19] Ali, Khawaja Haider, Marvin Sigalo, Saptarshi Das, Enrico Anderlini, Asif Ali Tahir, and Mohammad Abusara. 2021. "Reinforcement Learning for Energy-Storage Systems in Grid-Connected Microgrids: An Investigation of Online vs. Offline Implementation" Energies 14, no. 18: 5688. https://doi.org/10.3390/en14185688.

[20] A computationally Efficient Optimization Approach for Battery Systems in Islanded Mi-crogrid. Avijit Das, Zhen Ni. 2017, IEEE Transsaction on Smart Grid, pp. 1-11.

[21] Optimal generation scheduling of a microgrid. X. Wu, X. Wang, and Z. Bei. Berlin : IEEE, 2012. 3rd IEEE PES, Innovative Smart Grid Technologies Europe (ISGT Europe), pp. 1-7.