Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Power Flow Management in Multi-Source Electric Vehicle Charging Station

Arwa O. Erick*. Komla A. Folly. **

*Department of Electrical Engineering, University of Cape Town, Private bag, X3, Rondebosch, 7701, Cape Town, South Africa (Tel: +27-737-332-322; e-mail: arweri@myuct.ac.za).
** (e-mail:komla.folly@uct.ac.za)

Abstract: Grid-tied renewable energy sources (RES) with battery-behind-meter (BBM) architectures have successfully been used to ensure effective energy cooperation between the grid and RES-based microgrids. Such environments are quite stochastic, thus making power management very challenging. This paper presents the use of an asynchronous Q-learning in performing a power flow management task in a multi-source electric vehicle charging station with the integration of vehicle-to-microgrid technology. The power scheduling problem is first formulated as a Markov decision process. Asynchronous Q-learning is then used to solve it. The algorithm is tested with a typical charging station load profile over a 24-hour period and compared with a simple rule-based algorithm. Simulation results show that the proposed method is able to select a power schedule that reduces the energy cost with a better utilization of both the battery storage system and the vehicle to microgrid energy compared to the rule-based method.

Keywords: Charging station, electric vehicle, Q-learning, reinforcement learning, renewable energy

1. INTRODUCTION

It is estimated that the transportation sector causes at least 19% of CO_2 emissions. The substitution of internal combustion engine vehicles (ICEVs) with electric vehicles (EVs) has been proposed to reduce greenhouse gas and particulate emissions (World Economic Forum, 2018). Therefore, grid-tied RES with battery-behind-meter (BBM) architectures have been used (Gucin, Ince and Karaosmanoglu, 2015), with rare cases of vehicle-to-microgrid (V2M) technology.

Despite the growing attention in control strategies for V2G applications in grid tied EV charging stations, little attention has been paid on the energy scheduling for such arrangements. Wang *et al.,* (Wang *et al.*, 2018) modelled an optimal energy flow system for a building with V2G technology. In (Li *et al.*, 2014) a real-time simulation of energy control for an EV charging station under V2G, grid-to-vehicle (G2V) and vehicle-to-vehicle (V2V) modes are implemented. The above authors used general rule-based approaches that do not guarantee optimal results. A priority-based method has been used in (Abronzini *et al.*, 2016). Power management in such environments is quite challenging as the variables to be considered such as the load, the grid tariff and the RES generator output are all stochastic.

The use of reinforcement learning techniques in power systems scheduling has attracted significant research attention recently due to their ability to operate well in stochastic environments. The main advantage of RL techniques over other optimization methods such as linear search methods and swarm intelligence is that they can learn an optimal policy for a general load and generation profile and generate optimal solutions for online operations without the need to iterate every time a new load and generation profile is introduced (Kim and Lim, 2018). A more detailed description of reinforcement learning approaches to power systems scheduling and application of Q-learning in optimization can be found in (Jasmin, 2008), and (Kim and Lim, 2018).

There are two main Q-learning methods, namely, synchronous Q-learning and asynchronous Q-learning. In synchronous Q-learning, the agent sweeps through the entire state-space in every episode while in the asynchronous method, states are indexed and added to the state-action table as they occur, thus, it is a "sweepless" Q-learning technique (Volodymyr *et al.,* 2016). Synchronous Q-learning has been used in battery scheduling (Kim and Lim, 2018), (Kofinas and Dounis, 2018). However, synchronous Q-learning suffers from the curse of dimensionality and slow training when the Q-table approach is used (Jasmin, 2008). In asynchronous Q-learning, the states are added to the Q-table table as they occur and are accessed randomly (Travnik *et al*., 2018). Therefore, parallel learning can be employed to speed up training which may be challenging with synchronous methods where states are accessed sequentially. (Volodymyr *et al.*, 2016).

In (Arwa and Folly, 2020), a Q-learning algorithm has been applied to perform energy scheduling in a PV/battery EV charging station to minimize energy costs as well as maximizing the revenues from energy sales to the utility grid, but without the integration of V2M technology. In (Kuznetsova *et al.*, 2013), (Leo, Milton and Sibi, 2014), (Foruzan, Soh and Asgarpoor, 2018), (Xi *et al.*, 2015) and (Kim and Lim, 2018), Q-learning has been applied to schedule sources in grid-tied RES-based systems, all without the inclusion of V2M technology.

This paper uses a Q-table based asynchronous Q-learning technique to solve the power control problem in a multi-source EV charging station with vehicle-to-microgrid (V2M) integration. The power management problem for a multi-source EV charging station is first formulated as an MDP, then an asynchronous Q-learning algorithm in which both states and control actions are indexed in a hash-table structure is used to solve it. The algorithm decides when and how much power from each of the charging station's power sources is used to satisfy the charging load profile within the station's

constraints. A simple rule-based algorithm is then implemented and compared with the proposed algorithm.

2. THE EV CHARGING STATION

A grid-tied solar-powered EV fast-charging station (CS) with a battery storage system (BSS) and V2M integration is considered. A common DC bus is used to facilitate power-sharing among the electric vehicle supply equipment. The DC bus is linked to the grid through an AC to DC converter and the BSS and the V2M supply through a suitable DC to DC converter. The main role of the BSS is to store the energy of the solar PV during low grid tariff hours to be used during high tariff hours. Therefore, it functions as both a producer and a consumer of energy. The V2M power is supplied by vehicles whose owners have signed an energy supply contract with the station operators.

3. MATHEMATICAL FORMULATION

The main objective of this optimization is to minimize the operational cost of supplying a given EV charging load profile within the constraints of the charging station. The instantaneous power balance equation that guarantees that the load demand at the charging station is met is given by:

$$P_{cl}(t) = P_{bss}(t) + P_g(t) + P_{v2M}(t) + P_{pv}(t), \quad (1)$$

where P_{cl}, P_{pv}, P_{bss}, $P_g(t)$ and P_{v2M} are the station's charging load, the PV generator output, the battery charge/discharge power, the power purchased from the utility grid and the power purchased from the vehicle to microgrid (V2M) scheme. It is assumed that the grid does not absorb power from the charging station but only supplies power. Therefore, $P_g(t)$ is always positive. Also, $P_{bss}(t)$ is taken to be positive when it is supplying power to the load at the CS (discharge mode) and negative when it is taking power from the other sources (charging mode).

The total operational cost (C_{tot}) is the sum of the cost of power purchase from the grid $C_{P_g}(t)$, cost of power purchase from V2M $C_{P_{v2M}}(t)$ and battery degradation cost $C_{P_{bss}}(t)$.

The objective function of the system is, therefore, given by:

$$Min.(C_{tot}) = Min \sum_{t=0}^{T} \left[C_{P_g}(t) + C_{P_{v2M}}(t) + C_{P_{bss}}(t)\right]. \quad (2)$$

Equation (2) is subject to the constraints of power balance at the DC link given in (1), state of charge boundaries, grid power limits as dictated by a contract signed with the utility operators, and V2M energy limits agreed upon with the V2M vehicle owners.

Considering a real-time tariff system, the cost of power purchased from the grid at any time step is given by $C_{Pg}(t) = G_t(t)\, P_g(t)\Delta t$. The cost of power purchase from the V2M scheme is given by $C_{P_{v2M}}(t) = C_{fit}\, P_{v2M}(t)\Delta t$, where C_{fit} is the V2M feed-in tariff. This feed-in tariff is chosen so that the grid tariff shoots above it during peak load and goes below it during light load periods. The cost of drawing power from or storing power in the BSS is given by $C_{P_{bss}}(t) = P_{bss}(t)C_{bd}(t)\Delta t$, where $C_{bd}(t)$ is the cost of degradation of BSS given in equation (3). To model the cost of battery degradation the contributions of temperature, C_T, depth of discharge (DoD), C_{DoD}, and the average SoC, (SoC_{av}), C_{SoC},) are considered. The degradation cost is given by, $C_{bd} = max\{C_T, C_{DoD}, C_{SoC}\}$. The equations for each contribution have been derived by (Badawy and Sozer, 2017) as shown below.

$$C_{bd} = C_{bt} max\left\{\left(\int_{t_o}^{t_f} \frac{dt}{Y_h L_t(T)}\right), \left(\left[\left(\frac{1}{L(DoD_2)}\right) - \left(\frac{1}{L(DoD_1)}\right)\right]\right), \left(\frac{mSoC_{av} - d}{Q_{fade} n Y_h}\right)\right\} \quad (3)$$

In equation (3), C_{bt} is the battery capital cost per kWh, t_o and t_f are initial and final battery operation time for charge or discharge operations, Y_h is the number of hours in a year, $L(DoD_j)$ is the cycle life of the battery at DoD_j, $L_t(T)$ is the relationship of battery lifetime with temperature, Q_{fade} is the capacity fade at the battery end of life, SoC_{av} is the average SoC and m, n and d are curve fitting constants.

4. MDP MODEL OF THE POWER MANAGEMENT PROBLEM

In this section, the problem defined in section 3 is expressed as a Markov Decision Process (MDP) which is the formal mathematical construction for reinforcement learning. An MDP is defined by the tuple (*S, A, F, R)*, where *S* is the system state, *A* is the possible decision action, *F* is the state transition function and *R* is the reward obtained in taking the action in state *S*.

4.1 State Model

The state, x_k of the system is the set $\{k, P_{cl}^k, P_{pv}^k, G_t^k, E_b^k, E_{v2M}^k\}$, where k is the time component: $k = 0,1,\dots,T-1$, T is the optimization horizon, P_{cl}^k is the load at the CS at k, P_{pv}^k is the solar PV generation at k, G_t^k is the grid tariff at time k, all of which are forecasted. E_b^k and E_{v2M}^k are the amounts of energy in the BSS and the V2M battery packs respectively. The state-space χ is, therefore: $\chi = x_0 \cup x_1 \cup, \dots, \cup x_{T-1}$.

4.2 Action Model

The action vector at a time k is given by $a_k = \{P_{bss}^k, P_g^k, P_{v2M}^k\}$. The action space is a function of the state: $\mathcal{A}_k = f(x_k)$. Only actions that meet the system constraints are included in the set of possible actions so that $P_g^{min} \le P_g^k \le P_g^{max}$ and $E_b^{min} \le E_b^k - P_{bss}^k \Delta t \le E_b^{max}$ and $E_{v2M}^{min} \le E_{v2M}^k - P_{V2G}^k \Delta t \le E_{v2M}^{max}$.. The overall action space is therefore given by the union of all sets of individual state action spaces: $\mathcal{A} = \mathcal{A}_0 \cup \mathcal{A}_1 \cup, \dots, \cup \mathcal{A}_{T-1}$.

4.3 Model of State Transition

The state transition is defined as: $x_{k+1} = f(x_k, a_k)$, where x_{k+1} is the vector of the system inputs for the next state with elements of load P_{cl}^{k+1}, solar PV generation P_{pv}^{k+1}, the grid tariff G_t^{k+1} and the updated state of energy levels of the BSS and V2M batteries are given by $E_b^{k+1} = E_b^k \pm P_{bss}^k \Delta t$ and $E_{v2M}^{k+1} = E_{v2M}^k \pm P_{v2M}^k \Delta t$ respectively. Assuming that the forecasted values are correct, the state transition is given by,

$$x_{k+1} = \{k+1, P_{cl}^{k+1}, P_{pv}^{k+1}, G_t^{k+1}, E_b^{k+1}, E_{v2M}^{k+1}\}. \quad (4)$$

4.4 Agent's Reward Model

The reward, $r(k) = g(x_k, a_k, x_{k+1})$, is defined so that maximizing total reward leads to minimization of the global cost as given in equation (5) below.

$$r(k) = \frac{1}{C_{P_g}(t)+C_{P_{v2M}}(t)+C_{P_{bss}}(t)+1} \quad (5)$$

5. RULE-BASED ALGORITHM AND ASYNCHRONOUS Q-LEARNING

5.1 Rule-based Approach

A rule-based algorithm was implemented to solve the MDP developed in section 4. In this rule-based method, the rule was that for every time step, the state variables were considered, and a list of possible control actions that meet the load demand was produced. The instantaneous cost of each control action was then computed using equation (2) and the action that minimizes the instantaneous cost within the system constraints was selected. This was done across the whole state space.

5.2 Q-learning Background

In Q-learning, the agent learns the optimal policy represented by π^*, mapping every state to the optimal action. The value function, $Q(x, a)$ denotes how good it is to take an action a in state x, such that (Watkins and Dayan, 1992):

$$Q(x, a) = r_x(a) + \gamma \sum_y P_{xy}[\pi(x)]V^{\pi}(y) \quad (6)$$

where $r_x(a)$ is the immediate reward, γ the discount factor, P_{xy} the state transition probability, $V^{\pi}(y)$ is the value of the future state y visited by the agent as a result of taking action a in state x. In every learning episode, the agent visits each of the states x_k, with an action space, $\mathcal{A}_k$, selects an action a_k using the $\epsilon-$greedy method, and as a result, transits to the next state x_{k+1}, receiving an immediate reward, $r = g(x_k, a_k, x_{k+1})$. Then the Q-values are updated in the Bellman fashion as in equation (7) (Watkins and Dayan, 1992).

$$Q^{n+1}(x, a) = Q^n(x, a) + \alpha\big[r(k) + \gamma max_{a_{k+1}} Q^n(x_{k+1}, a_{k+1}) - Q^n(x, a)\big] \quad (7)$$

where $\alpha \in (0,1)$ is the learning rate that determines the extent of modification of Q-values, $Q^n(x, a)$ is the current Q-value, $Q^{n+1}(x, a)$ is the next Q-value while $\gamma \in (0,1)$ is the discount factor. If α is sufficiently small, Q^n converges to Q^* after enough iterations. If the current state is terminal, then there is no next state, hence the Q-value update is given as follows:

$$Q^{n+1}(x, a) = Q^n(x, a) + \alpha[r(k) - Q^n(x, a)]. \quad (8)$$

5.3 Proposed Asynchronous Q-learning Algorithm

In the proposed asynchronous Q-learning, the Q-table is created as an empty dictionary (hash table) into which states are added as keys, with dictionaries of the allowable actions and their initial Q-values, as values. The Q-table, therefore, becomes a nested dictionary, with state indices as keys. And for every state index, there is a dictionary of possible actions, with actions as keys and corresponding Q-values as values. The states and their possible actions are added as they occur during learning to keep the algorithm more computationally efficient and fast. The modified Q-learning algorithm is given in Fig. 1.

6. SIMULATION RESULTS AND DISCUSSIONS

6.1 Simulation

A typical grid-tied solar-powered fast EV charging station has been considered with a limited supply of solar power to supply the load profile in time steps of 1hour for a 24-hour optimization horizon. For the purpose of this simulation, it is assumed that the grid and the V2M battery packs only supply power deficit and do not absorb any power from the station.

```
1.  Initialize learning parameters (α and γ)
2.  Initialize epsilon (ϵ = 1)
3.  Create an empty q-table
4.  For n = 0 to maximum iteration:
5.      k = 0
6.      Get the initial state vector
7.      For k = 0 to T-1:
8.          If the state is not in q-table:
9.              Get possible actions
10.             Initialize q-values = 0
11.             Add state-action pair to q-table
12.         End if
13.         Take action by ϵ −greedy method
14.         Find next state
15.         Calculate reward
16.         Update q-values
17.         k = k + 1
18.     End for
19.     Update ϵ
20.     n = n + 1
21. End for
22. Return q-table
```

Fig. 1: Asynchronous Q-learning Algorithm

Similarly, a typical day-ahead dynamic tariff forecast curve has been chosen with energy cost per kWh given in USD. The inputs to the algorithm include the forecasted PV generation profile, the day ahead grid tariff profile and the load profile at the CS as given in Fig. 2.

Table I shows the selected Q-learning hyperparameters. The initial exploration rate, ϵ, was chosen to be 1 to ensure maximum exploration of the action space during initial episodes of learning. A discount factor of 1 was also selected as the future costs are just as significant as immediate costs. This ensures that control actions are valued for both their current effect and the effect they would have in the proceeding states. The learning rate, α, was selected by trial and error, and a value of 0.001 was found to give the best convergence.

Table 1. Learning Hyperparameters

Hyperparameter	Chosen value
Initial exploration rate (ϵ)	1.0
Learning rate (α)	0.001
Discount factor (γ)	1.0

The simulation parameters are given in Table 2. The grid limits are subject to a contract signed by the station owners with the grid operators and must never be exceeded. Also, limiting the maximum grid power to just the maximum expected deficit helps reduce the action space. The BSS capital cost and temperature characteristics have been taken from (Badawy and Sozer, 2017). The Q-learning algorithm is then implemented in Python programming language.

Table 2. The charging station simulation parameters

Parameter	*Symbol*	*Values*
Min./Max. grid power	P_g^{min}/P_g^{max}	0.0/55kW
Timestep	Δt	1hour
Grid feed-in tariff	C_{fit}	\$0.3/kWh
Battery capital cost	C_{bt}	\$400
Battery capacity	E_b	100kWh
Initial battery energy	SoE_{in}	40kWh
Min. battery energy	E_b^{min}	20kWh
Min./max. V2M Battery Energy	$E_{V2M}^{min}/E_{V2M}^{max}$	20/60 kWh
BSS thermal resistance	R_t	0.2 mΩ

6.2 Results and Discussions

Fig. 2 shows the grid power schedule alongside the load, PV and grid tariff profiles obtained using the proposed algorithm. It can be seen that the proposed algorithm reduces grid power purchase during peak tariff and maximizes the use of grid power to supply the load during low tariffs. The dynamic tariff is a very good indicator of the load profile on the utility grid as in most cases, it rises and falls with the load demand. Therefore, as the grid tariff is hiked during high load demand on the utility grid, the algorithm's decision to minimize the power purchase from the grid not only helps the station reduce the energy costs but also reduces the strain on the grid. However, for a situation where the load is high and PV is insufficient, the grid power purchase may be scheduled even though the grid tariff is expensive, as can be seen between the 5th and the 10th hours.

Fig. 3 shows the BSS state of energy schedule obtained from the proposed method plotted alongside the load at the CS, the grid tariff and the PV output profiles. In the beginning, the battery discharges to supply the load, but in the 2nd hour, charging begins in preparation for the tariff hike in the proceeding hours. It can be seen that the battery energy drops significantly when the tariff peaks at the 6th hour when and PV is insufficient to supply the load. Further attempt to charge is limited by the charging station load demand between the 10th and the 15th hours and low PV generation between the 15th and the 24th hour. In this setup, the PV output was always insufficient, thus, the battery generally showed a discharge trend.

Fig. 4 shows the utilization of the V2M energy over the 24 hours plotted with the load, the tariff, and the PV output obtained using the proposed method. For simplicity, it is assumed in this work that the V2M vehicles do not absorb energy from the charging station, therefore, the algorithm's action is limited to deciding on when and how much power should be drawn from it. Therefore, cost-saving is in using its available energy at times when the grid is expensive and the charging station faces a high demand. There is low usage of the V2M energy when the load demand at the CS is low and the grid tariff is lower than the V2M feed-in tariff as seen between 0 and 5th hour to preserve the V2M energy for use during high grid tariff periods. This is followed by a sharp discharge of the V2M battery between the 5th and the 6th hours when the station load and the grid tariff peaks above the V2M feed-in tariff as it is more economical to use the cheaper V2M power at that moment.

In Fig. 5, the V2M energy schedule obtained by the proposed method is compared with the one obtained from the rule-based algorithm. As mentioned previously, the V2M is only allowed to supply energy to the CS. It can be seen that the rule-based method sharply discharges the V2M battery immediately after the 4th hour when the grid tariff goes above the V2M feed-in tariff of 0.3 USD/kWh. Although this control action is cost-effective in the short term, in the long-term, it is cheaper to delay this energy uptake until the grid tariff peaks to its maximum value of 0.5 USD/kWh. The proposed method was able to achieve this long-term goal of reducing the cost of electricity by causing the V2M to discharge between the 5th and the 6th hour when the grid tariff was rising to its peak value

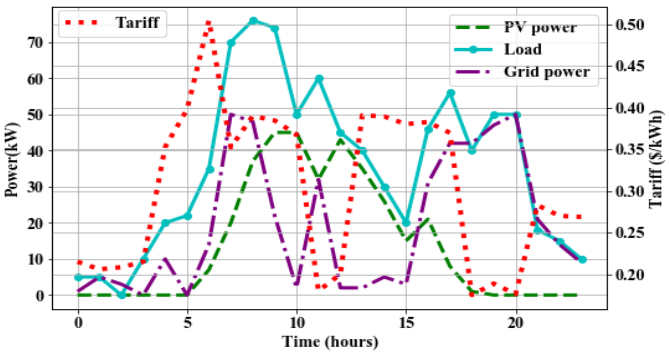

Fig. 2: Grid power, battery power and PV power over 24 hours

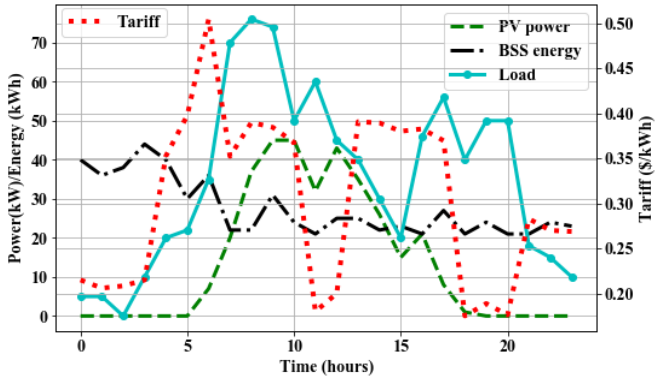

Fig. 3: Battery Energy schedule, day-ahead tariff, load and PV profiles

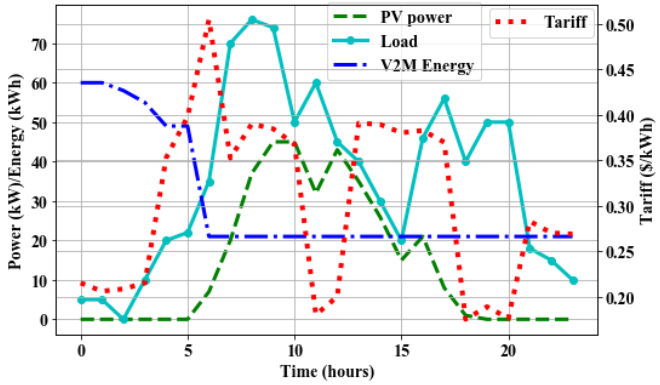

Fig. 4: V2M battery Energy schedule alongside tariff, load and PV profiles

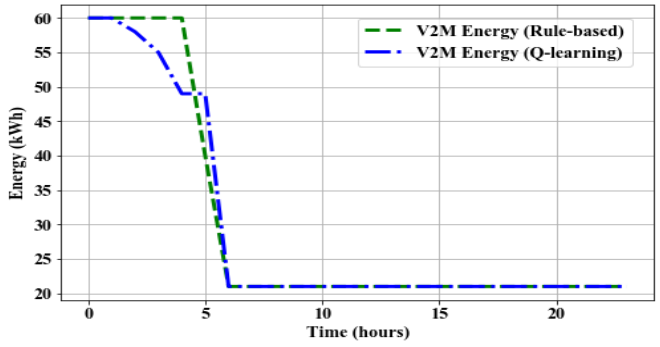

Fig. 5: V2M energy schedule for the proposed Q-learning and rule-based methods alongside grid tariff

In Fig. 6, the BSS energy schedule obtained by the proposed asynchronous Q-learning is presented alongside the schedule obtained by the rule-based method. It can be seen that the rule-based algorithm discharges the BSS all the time except when the load at the CS is zero as at the 2nd hour as shown in Fig. 2. As a result, the battery is drained to its minimum value of 20kWh quickly. Although this control action is cheaper in the short-term, it is expensive in the long term as the battery gets completely drained, thus the CS is left with only the grid power and the PV, so that if the PV output drops below the load demand, the CS will have to use the expensive grid power to supply its load. In comparison, the proposed method performs proactive charge schedules when the grid tariff is cheap and discharges the BSS when the grid tariff is expensive, thus reducing the cost of power purchase from the grid. As such, this algorithm effectively uses the BSS to obtain a schedule with a lower cost.

Fig. 7 shows the cost profile for the selected episode with the load and PV generation profiles. Energy cost is highest during high grid tariff, high load, and low PV output, and drops significantly as PV generation increases. This shows that the proposed method ensures maximum self-consumption of the station generated PV power which effectively reduces the cost of energy. As a result of this schedule obtained from the proposed method, a global cost of 143 USD was returned as compared to 148 USD returned by the rule-based method.

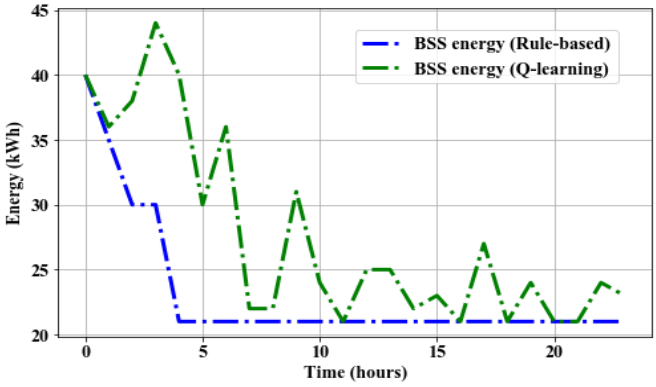

Fig. 6: BSS Energy Schedule for the two algorithms

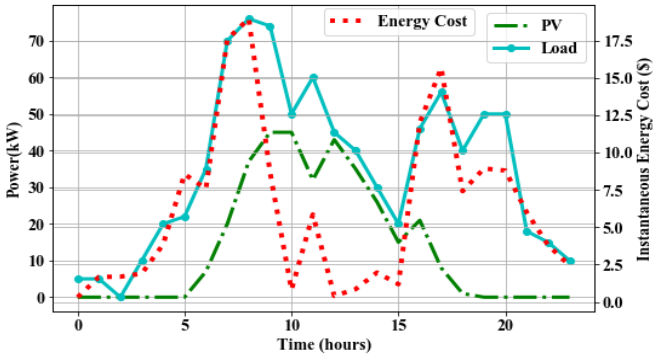

Fig. 7: Energy Cost profile for optimal episode with PV and load

Fig. 8 shows the total running cost variations during training. The resultant schedule (around episode 140000) returned a cost of as low as 143 USD as opposed to earlier episodes (around episode 20000) that had a cost as high as 190 USD. The final global cost using the proposed method is 143 USD while the one obtained through the rule-based method is 148 USD. Though this difference is small when considering only one day, adequate cost savings can be obtained over the years. The proposed method returned a lower global cost despite having performed much more charge-discharge cycles on the BSS which should have led to a higher global cost due to higher battery degradation cost incurred.

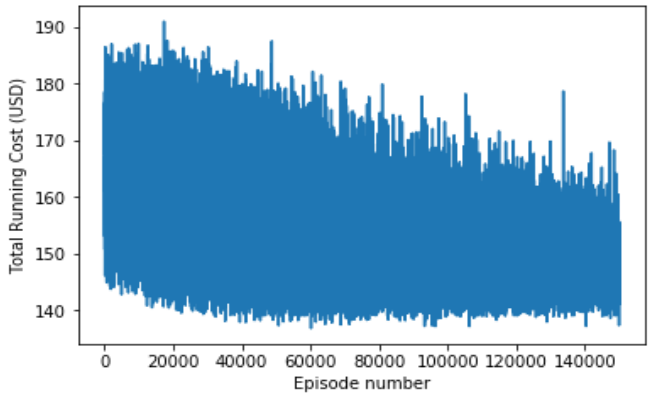

Figure 8: The learning curve for the Q-learning algorithm

6. CONCLUSION

In this paper, the power management problem in a grid-tied PV/battery EV charging station with the integration of V2M has been defined as an MDP and solved using asynchronous Q-learning. It can be concluded that the proposed method returned a lower global cost and better battery utilization than a rule-based method. The advantage of using the asynchronous method is that learning can be parallelized to speed up the learning process. However, the asynchronous Q-learning method is very sensitive to the learning hyperparameters which are difficult to tune in order to get a proper convergence. Also, Q-learning employing a Q-table, regardless of the design, still suffers from the curse of dimensionality which severely affects its robustness. In the future publication, a deep reinforcement learning technique that is more robust will be considered to solve the scheduling problem.

REFERENCES

U. Abronzini, C. Attaianese, M. D'Arpino, M. Di Monaco, A. Genovese, G. Pede, G. Tomasso (2016) 'Optimal energy control for smart charging infrastructures with ESS and REG', *2016 International Conference on Electrical Systems for Aircraft, Railway, Ship Propulsion and Road Vehicles and International Transportation Electrification Conference, ESARS-ITEC 2016*. IEEE, pp. 1–6. DOI: 10.1109/ESARS-ITEC.2016.7841427.

Arwa O. Erick; Komla A. Folly (2020) 'Energy Trading in Grid-connected PV-Battery Electric Vehicle Charging Station', *Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*.

Badawy, M. O. and Sozer, Y. (2017) 'Power Flow Management of a Grid-Tied PV-Battery System for Electric Vehicles Charging', *IEEE Proceedings of the 2017 Winter Simulation*, 53(2), pp. 1347–1357. DOI: 10.1109/TIA.2016.2633526.

Foruzan, E., Soh, L. K. and Asgarpoor, S. (2018) 'Reinforcement Learning Approach for Optimal Distributed Energy Management in a Microgrid', *IEEE Transactions on Power Systems*. IEEE, 33(5), pp. 5749–5758. DOI: 10.1109/TPWRS.2018.2823641.

Gucin, T. N., Ince, K. and Karaosmanoglu, F. (2015) 'Design and power management of a grid-connected Dc charging station for electric vehicles using solar and wind power', *2015 3rd International Istanbul Smart Grid Congress and Fair, ICSG 2015*, (May). DOI: 10.1109/SGCF.2015.7354921.

Jasmin, E. A. (2008) 'Reinforcement Learning Approaches to Power System Scheduling'. Ph.D. Thesis, School of Engineering, Cochin University of Technology.

Kim, S. and Lim, H. (2018) 'Reinforcement Learning Based Energy Management Algorithm for Smart Energy Buildings', *Energies*, 11(8), p. 2010. DOI: 10.3390/en11082010.

Kofinas, P., Vouros, G. and Dounis, A. I. (2018) 'Energy management in solar microgrid via reinforcement learning using fuzzy reward', *Advances in Building Energy Research*, 12(1), pp. 97–115. DOI: 10.1080/17512549.2017.1314832.

Kuznetsova, E. Li, F., Ruiz, C., Zio, E., Ault, G., and Bell, K., (2013) 'Reinforcement learning for microgrid energy management', *Energy*. Elsevier Ltd, 59, pp. 133–146. DOI: 10.1016/j.energy.2013.05.060.

Leo, R., Milton, R. S. and Sibi, S. (2014) 'Reinforcement learning for optimal energy management of a solar microgrid', *2014 IEEE Global Humanitarian Technology Conference - South Asia Satellite, GHTC-SAS 2014*. IEEE, pp. 183–188. DOI: 10.1109/GHTC-SAS.2014.6967580.

Shuhui Li, Ke Bao, Xingang Fu and Huiying Zheng (2014) 'Energy management and control of electric vehicle charging stations', *Electric Power Components and Systems*, 42(3–4), pp. 339–347. DOI: 10.1080/15325008.2013.837120.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Tim Harley, Timothy P. Lillicrap, David Silver, K. K. (2016) 'Asynchronous Methods for Deep Reinforcement Learning', *JMLR: W&CP, Proceedings of the 33rd International Conference on Machine Learning, New York, USA*, 48. Available at: http://arxiv.org/abs/1301.3781.

Wang, Z., Tang, Y., Chen, X., Men, X., Cao, J. and Wang, H. (2018) 'Optimized daily dispatching strategy of building-integrated energy systems considering vehicle to grid technology and room temperature control', *Energies*, 11(5). DOI: 10.3390/en11051287.

Watkins, C. J. C. H. and Dayan, P. (1992) 'Technical Note: Q-Learning', *Machine Learning*, 8(3), pp. 279–292. DOI: 10.1023/A:1022676722315.

World Economic Forum (2018) 'Electric Vehicles for Smarter Cities: The Future of Energy and Mobility', *World Economic Forum*, (January), p. 32. Available at: http://www3.weforum.org/docs/WEF_2018_Electric_For_Smarter_Cities.pdf.

Yuanyuan Xi, Liuchen Chang, Meiqin Mao1, Peng Jin, Nikos Hatziargyriou, Haibo Xu (2015) 'Q-learning algorithm based multi-agent coordinated control method for microgrids', *9th International Conference on Power Electronics - ECCE Asia: 'Green World with Power Electronics', ICPE 2015-ECCE Asia*. Korean Institute of Power Electronics, pp. 1497–1504. doi: 10.1109/ICPE.2015.7167977.