Demand Response using Electric Vehicles in Mutli-Agent Multi-Objective Cooperative Game

Vadim Avkhimenia
Electrical and Computer Engineering
University of Alberta
Edmonton, Canada
avkhimen@ualberta.ca

***Abstract* – Electric vehicles (EVs) vehicle-to-grid (V2G) demand response involves charging and discharging EVs to aid in the efficient grid operation. We attempt to introduce a V2G scheme based on EV state-of-charge (SOC) regions cooperatively acting as agents in a multi-agent, multi-objective game to minimize the daily Peak-to-Average ratio (PAR) and maintain an appropriate SOC level by the morning. It was found that this scheme does not reduce PAR but results in a 2.7% average reduction in demand load standard deviation for the full day cycle, with higher reduction of 16.2% for the first 23 hours of the day.**

Keywords—Electric Vehicles, vehicle-to-grid, V2G, EV, Reinforcement Learning, Multi-Agent, Multi-Objective, MARL, Cooperative Game, Demand Response, Peak-to-Average Ratio

I. INTRODUCTION

The transition towards renewable energy is driven by governments via financial incentives and policy. International Energy Agency (IEA) proposed the transition towards the electrification of transportation towards 2030 through the Stated Policies Scenario, and the Sustainable Development Scenario. According to the Sustainable Development Scenario, 30% of cars globally should be converted to electric by 2030 [1]. Under the Sustainable Development Scenario, the stock of electric vehicles worldwide will need to grow by 50 times by 2030. The realistic forecasts from the IEA estimate an annual growth of 30% [2].

All these new electric vehicles will need to be charged. Right now, most electric vehicles begin charging around 5-6 PM which creates a high load spike on the grid. In the case of 10% market penetration of EVs an increase in the peak load demand is estimated to be around 18%, and much higher for higher levels of market penetration [3], [4]. EV batteries can also be used as sources of power by discharging them. Under the Sustainable Development Scenario, the amount of potential power EVs can provide into the grid can be substantial [5]. Benefits of EVs are peak load shaving, spinning reserve provision, load management, stabilization of grid voltage, line loss reduction, and frequency control [6], and the negative impacts of EVs are voltage instability, increased peak demand, power quality problems, increased power loss, and transformer overloading [7].

Electrical grid is designed to satisfy peak demand of users. This means installing new expensive generation and transmission equipment which can put a strain on the system [8]. This is mitigated via demand response schemes aimed at reducing PAR, which is the ratio of the maximum load to the average load over a time period [9]. One effective scheme is to use charging and discharging of electric vehicles (EVs) to control the level of energy in the grid [10].

My objective is to reduce PAR via effective charging and discharging of EVs using a new proposed V2G scheme based on EVs with similar SOC levels acting cooperatively to reduce PAR. In section II relevant works will be presented. In section III the system model will be described, in section IV the problem will be formulated, and in section V the problem solution will be presented. A case study will be shown in section VI and the results will be discussed in section VII.

II. RELATED WORKS

In [11] the design of an intelligent socket outlet that utilized an algorithm using real-time power demand, the EV's state-of-charge (SOC), and user charging preference was developed. Algorithm was successfully able to reduce the demand peak by shifting charging to off-peak hours. Based on demand data, EV battery specs, SOC of EV, and required completion time, the method calculates whether to charge the EV in QUICK, BUDGET, or CLEVER mode and when to begin charging. [9] proposed an optimization technique which calculates scheduling vectors for charging and discharging of EVs by minimizing objective functions of demand, energy cost, and time of charging/discharging. Algorithm successfully reduces PAR, decreases charging cost, and increases discharging cost. Multi-objective linear programming method (MOLP) is used to run the algorithm to solve for the scheduling vectors until the algorithm converges. [12] proposed a decision-tree-based algorithm to reduce residential load using EVs, photovoltaic panels (PVs), and battery energy-storage systems (BESSs). By peak-shaving, the algorithm reduced peak residential load by 53%. EV were modelled to discharge only when their SOC level was above specified minimum discharge threshold. [13] proposed a 2-stage peak-shaving vehicle-to-grid (V2G) scheme to minimize the peak demand utilizing EV discharging. The algorithm was able to reduce the peak demand with a peak-shaving index of 98% at 20% EV

penetration. the peak shaving level was determined based on the load profile. In [14] a real-time water-filling algorithm was proposed that can be integrated with commercial buildings to reduce peak demand by peak shaving. With priority charging the algorithm managed to reduce the demand charge by 80% of theoretical maximum. Optimization problem between the charging rates and the remaining energy to be charged was solved and charging rate for each EV was outputted. In [15] a charging framework was developed for the real-time dispatches of EVs via minimizing the load variance and reducing the cost of charging. Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) approach managed to minimize losses for the 80% EV penetration case. A multi-objective model to minimize the charging cost, battery degradation, and load variance was setup. [16] proposed a decision-tree-based algorithm to reduce commercial load using EVs, PVs, and BESSs. The algorithm reduced peak industrial load by 50%. Decisions were made by comparing reference shaving point with the building energy demand. [17] modelled the behavior of multiple EV charging stations with the objective to minimize the cost of each charging station via the quadratic optimization function. The master agent was modelled to send the load and voltage schedule for next day to each agent which used quadratic optimization to calculate their charging preferences, which are then sent back to the master agent for acceptance. The method allows for the voltage profile of each feeder under study to remain within acceptable limits and shifted the peak demand to low-demand hours. In [18] Q-Learning was used to minimize the energy loss of the hybrid energy storage system and the resulting method was shown to have potential to improve efficiency and reduce the energy loss. In [19] a system modelling the performance of a single agent EV charging station was proposed. The system aimed to maximize the revenue of the EV charging station. The systems utilized Q-Learning and achieved an increase in revenue of 40-80% demonstrating that the approach can be extended. In [20] the minimization cost of EVs and the degradation of the charging station was considered while taking into account user dissatisfaction. To model the problem suboptimal charging algorithm with constraints (SCAC) based on Lyapunov optimization was used. SCAC was extended using Q-Learning (SCRL) by picking the optimal action using the criteria from the SCAC algorithm. The simulation involved 30-80 EVs and the CSRL algorithm was more efficient than SCAC algorithm.

III. SYSTEM MODEL

The proposed scheme assumes that Alberta has constant number of EVs, the base energy demand profile changes hour to hour but is the same for every day of the year [22], there is always enough capacity to supply any amount of power into V2G service, all EVs are the same and the presence of EVs does not alter Alberta base load, and, on hourly basis, and every EV is enrolled in V2G scheme but only some EVs participate. The Alberta demand load is shown in Fig. 1. Either 25%, 50%, or 75% of EVs participate during each hour. The participation level is known at the start of every hour. The V2G service starts at 5 PM when the SOC of every EV is 30% on average, with 10% standard deviation. Between 5 PM and 3 AM EVs can charge, discharge, or do nothing, provided they participate in V2G service. If an EV does not participate, it will drive 5 kms on average, with 5 kms standard deviation [9], [20]. Starting at 3 AM, 75% of EVs participate in V2G service. V2G service ends at 8 AM.

IV. PROBLEM FORMULATION

The problem is formulated such that the following question can be answered: what is the best possible action (charge, discharge, or do nothing) that each EV should do to minimize the daily PAR given its SOC, participation level, and current hour? Or, as an example: if its 7 PM, EV is parked, percent participation is 50%, and EV has 35% charge, should it charge, discharge, or do nothing?

A. EV SOC

Day is partitioned into N_H slots, $h \in [1, 2, 3, \dots, N_H]$. There is a total of N_{EV} EVs. For every time slot h, it is assumed the SOC of each EV $i \in [1, 2, 3, \dots, N_{EV}]$, $S_{i,h}$, must be between minimum SOC, S_{min}, and maximum SOC, S_{max}:

$$S_{min} \leq S_{i,h} \leq S_{max} \tag{1}$$

At the time slot when V2G service starts, $h_{V2G\ start}$, initial SOC of each EV i, is sampled from a normal distribution with mean $\mu_{SOC\ V2G\ start}$, and standard deviation $\sigma_{SOC\ V2G\ start}$ [13]:

$$S_{i,h} \sim min(|N(\mu_{SOC\ V2G\ start}, \sigma_{SOC\ V2G\ start}, N_{EV})|, S_{max}) \tag{2}$$

B. EV Driving

Participation of each EV i in V2G service is governed by EV participation status $X_{i,h}$, which is 1 if EV is participating, and 0 otherwise, with the probability of participation given by participation level Y_h:

$$X_{i,h} = \begin{cases} 0 & i \text{ is not participating, } P(0) = 1 - Y_h \\ 1 & i \text{ is participating, } P(1) = Y_h \end{cases} \tag{3}$$

The participation level, Y_h, can be one of 25%, 50%, or 75%:

$$Y_h \in \{25\%, 50\%, 75\%\} \tag{4}$$

If $S_{i,h}$ is less than or equal to S_{min}, EV cannot drive, otherwise EV will drive the distance assigned to it by the driving distance $d_{i,h}$, which is sampled from a normal distribution with mean $\mu_{driving}$, and standard deviation $\sigma_{driving}$ [13]:

$$d_{i,h} \sim |N(\mu_{driving}, \sigma_{driving}, N_{EV})| \tag{5}$$

Provided $X_{i,h} = 1$, EV battery discharges by multiplying the distance driven by EV i with EV driving discharge rate per km, $R_{driving\ discharging}$:

$$S_{i,h+1} = S_{i,h} - X_{i,h} d_{i,h} R_{driving\ discharging} \quad (6)$$

C. EV Charging

The SOC of EV i after charging in time slot h, is given by adding constant charging rate per hour $R_{battery\ charging}$ to current SOC up to a maximum SOC level, S_{max}:

$$S_{i,\,h+1} = min\left(S_{i,h} + R_{battery\ charging}, S_{max}\right) \quad (7)$$

Total charging load in time slot h, $e_{h,\ charging}$, is the sum of the constant charging rate drawn from the grid for each EV i that is being charged, $R_{grid\ charging}$, provided EV i is participating in V2G service, and its SOC is less than S_{max}:

$$e_{h,\ charging} = \sum_{i=1}^{N_{EV}} E\left[R_{grid\ charging} | X_{i,h} = 1, S_{i,\,h} < S_{max}\right] \quad (8)$$

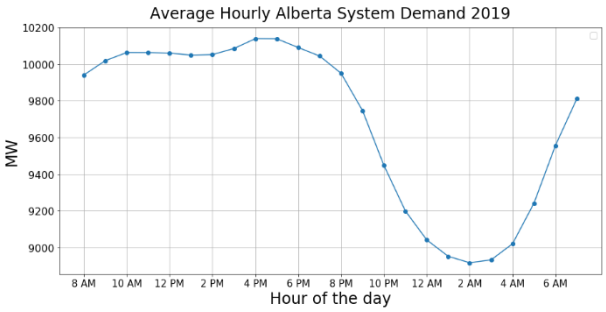

Fig. 1. Alberta daily average demand load for 2019

D. EV Discharging

The SOC of EV i after discharging in time slot h, is given by subtracting constant discharging rate per time slot $R_{battery\ discharging}$ from current SOC up to a minimum SOC level, S_{min}:

$$S_{i,\,h+1} = max\left(S_{i,h} - R_{battery\ discharging}, S_{min}\right) \quad (9)$$

Total discharging load in time period h, $e_{h,\ discharging}$, is the sum of the constant discharging rate fed into the grid for each EV i that is being discharged, $R_{grid\ discharging}$, provided EV i is participating in V2G service, and its SOC is greater than S_{min}:

$$e_{h,\ discharging} = \sum_{i=1}^{N_{EV}} E\left[R_{grid\ discharging} | X_{i,h} = 1, S_{i,\,h} < S_{max}\right] \quad (10)$$

The total hourly load, $e_{h,\ total}$, in time slot h, is then the difference between the charging and discharging EV loads, added with the Alberta base load, $e_{h,\ base\ demand}$ [3]:

$$e_{h,\ total} = e_{h,\ base\ demand} + e_{h,\ charging} - e_{h,\ discharging} \quad (11)$$

E. Objective

The problem is a minimization problem, where the quantity minimized is PAR for the number of time slots, N_H.

PAR for each time slot h is calculated by dividing the maximum load up to and include the time slot h, by average load up to and including the time slot h [9]:

$$PAR_h = h \frac{max\left(e_{h\in[1,h],\ total}\right)}{\sum_1^h e_{h,\ total}} \quad (12)$$

Minimum mean SOC for EVs at the time slot when V2G service ends, $h_{V2G\ end}$, must be at least $S_{V2G\ end\ min}$. The participation level is 75% starting at the time slot $h_{part\ 75\%\ start}$.

EVs act cooperatively to reduce the daily PAR, and at any time slot h, the SOC of each EV falls into one of $N_{SOC\ bins}$ SOC ranges. We assume 4 SOC ranges, 0-25%, 25-50%, 50-75%, or 75-100%.

The problem is then a multi-objective, multi-agent, cooperative game, with aim of each agent to minimize daily PAR while ensuring that average SOC of EVs at $h_{V2G\ end}$ is at least $S_{V2G\ end\ min}$.

For each hour, the actions of driving, charging, discharging, or doing nothing may cause some EVs to move into different SOC ranges.

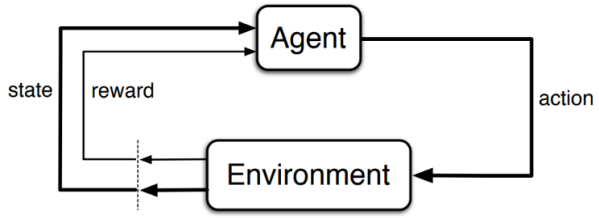

Fig. 2. Traditional discounted setting in reinforcement learning

V. PROBLEM SOLUTION

We transform the original multi-objective, multi-agent, cooperative game minimization problem, into a Markov Decision Process temporal-difference learning problem. In traditional reinforcement learning, the agent is in a state S, agent takes action A, environment puts agent into new state S', and provides the agent with reward R. Cycle repeats until the agent learns best actions for each state. On every step of the cycle there is a state-action value $Q(S, A)$ that maximizes future expected reward for the agent. The goal of this process is to find an action for each state that will maximize the sum of future rewards over all cycles, as illustrated in Fig. 2 [23].

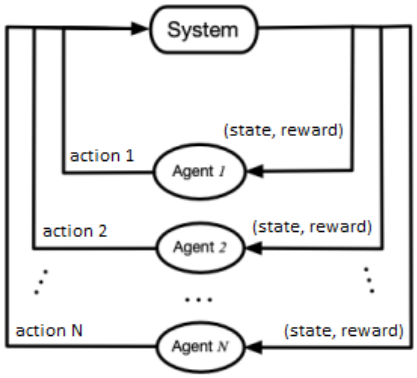

Fig. 3. Multi-agent cooperative game

In multi-agent cooperative reinforcement learning all agents are striving to maximize the same objective and on every transition are thrown into the same state and receive the same reward [24], [25], as illustrated in Fig. 3. When there two objectives, O_1, and O_2, they can be combined into one objective by linearly weighting them with weights w_1 and w_2 [26], [27]:

$$reward = O_1 w_1 + O_2 w_2 \tag{13}$$

A minimization problem is a maximization of the negative of the objective being studied.

We treat each SOC range as an agent, with all agents trying to minimize the daily PAR while ensuring that that mean SOC of EVs at the time slot $h_{V2G\ end}$ is at least $S_{V2G\ end\ min}$ [17], [21], [28].

The state is then defined as the combination of time slot h and percent participation, Y_h [19], [20]:

$$S = (h, Y_h) \tag{14}$$

The action, A_h, is to charge, discharge, or do nothing, selected randomly in each time period h:

$$A_h \in \{charge, discharge, do\ nothing\} \tag{15}$$

The penalty for the average SOC of EVs for the time slot h, P_h, is given as 3 if the mean SOC of EVs can reach at least the minimum required level by the end of the V2G service, and the penalty is given as -1, if it can not. If the time slot h is outside of the time period between when only 75% participation level is allowed and when V2G service ends, the penalty is defined to be 0. This choice of penalty encourages the agents to charge the SOC of each EV such that the mean SOC of EVs by $h_{V2G\ end}$ is at least $S_{V2G\ end\ min}$:

$$P_h = \begin{cases} 3 & S_{i,h} \geq B, \mathrm{h}_{\text{part 75\% start}} \leq h < \mathrm{h}_{\text{V2G end}} \\ -1 & S_{i,h} < B, \mathrm{h}_{\text{part 75\% start}} \leq h < \mathrm{h}_{\text{V2G end}} \\ 0 & h \notin [\mathrm{h}_{\text{part 75\% start}}, \mathrm{h}_{\text{V2G end}}] \end{cases} \tag{16}$$

where $B = S_{\text{V2G end}\,min} - (h_{V2G\ end} - h)R_{battery\ charging}$

The reward for the time slot h is linear combination of the negative PAR (to minimize) and the penalty for the average SOC to be above $S_{V2G\ end\ min}$ when V2G service ends:

$$r_h = -PAR_h + P_h \tag{17}$$

To solve this problem, we utilize Q-Learning algorithm. Q-Learning is an off-policy, model-free learning algorithm [23] which ensures convergence provided each state and action are visited an infinite number of times with decreasing learning rate.

Algorithm 1 Q-Learning

Input: learning rate $\alpha = [0,1]$
Input: exploration parameter $\epsilon < 0$
Init.: $Q(s, a)$, for all s, a, except $Q(terminal, \_) = 0$
For each episode:
  Get S
  For each step:
    Choose A from S using behavioral policy
    Take action A, observe R, S'
    $Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma\, max_{a'}(S', a') - Q(S,A)]$
    $S \leftarrow S'$

The optimum policy, π^*, from Q-Learning is obtained by greedyfying the state-action value function $Q(S,A)$ with respect to the action. This will produce the optimum action in each state.

Combining Q-learning, with the definitions of the state, action, reward, and environment, the general proposed algorithm is then:

Algorithm 2 Proposed Algorithm

Input: learning rate $\alpha = [0,1]$
Input: exploration parameter $\epsilon < 0$
Init.: $Q[(h, Y_h), a]$, for all h, Y_h, a, except $Q[terminal, \_] = 0$
For each episode:
  For each $h \in [1, 2, 3, \ldots, N_H]$**:**
    Choose Y_{h+1} at random
    For each SOC bin $[1, 2, 3, \ldots, N_{SOC\ bins}]$:
      Choose action, A, using ε-greedy
      Take chosen action A
    Calculate $e_{h,\ total}$
    Observe reward r_h and next state $(h+1, Y_{h+1})$
    For each SOC bin $[1, 2, 3, \ldots, N_{SOC\ bins}]$:
      $Q[(h, Y_h), A] \leftarrow Q[(h, Y_h), A] + \alpha(r_h + max_{a'}(Q[(h+1, Y_{h+1}), a']) - Q[(h, Y_h), A])$
  $(h, Y_h) \leftarrow (h+1, Y_{h+1})$
  Until $h = \mathrm{h}_{\text{V2G end}}$

VI. CASE STUDY

Day is partitioned into 24 time slots. We use the parameters shown in tables I and II [9], [19], [20]. To ensure adequate

time to produce the results the model was scaled so the ratio of maximum peak to maximum possible charging load from EVs is equal in scaled and un-scaled simulations. The simulation was repeated for different levels of market penetration with the number of EVs, N_{EV} = 100k, 200k, 300k, 400k, 500k, 600k, 700k, and 800k EVs, and scale 1000.

TABLE I. V2G SIMULATION PARAMETERS

V2G Simulation Parameters	
Parameter	***Value***
Minimum EV SOC	0.1
Maximum EV SOC	1
EV driving discharge rate per km	0.0035/hour
EV battery charging rate	0.15/hour
EV battery discharging rate	0.15/hour
Mean SOC of EVs at 5 PM	0.3
Standard deviation of EV SOCs at 5 PM	0.1
Minimum mean SOC of EVs at 8 AM	0.48
Mean driving distance in each hour	5km
Standard deviation of driving distance in each hour	5km
EV charging energy drawn from grid	10kW
EV discharging energy fed into grid	10kW

TABLE II. SIMULATION HYPER-PARAMETERS

Simulation Hyper-Parameters	
Hyper-Parameter	***Value***
Learning rate, α	0.01
Discount factor, γ, (episodic task)	1
Exploration parameter, ϵ	0.1

Fig. 4-7 were obtained for training of reducing PAR as a function of the number of training epochs for 200k EVs trained for 13,000 epochs over a single run and averaged over 100 steps.

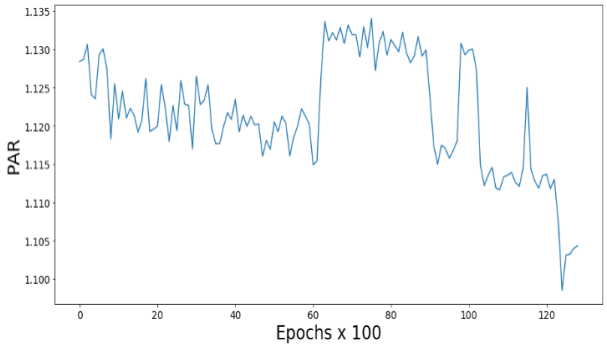

Fig. 4. PAR vs Epochs for 200k EVs computed over a single run of 13,000 epochs and averaged over 100 steps

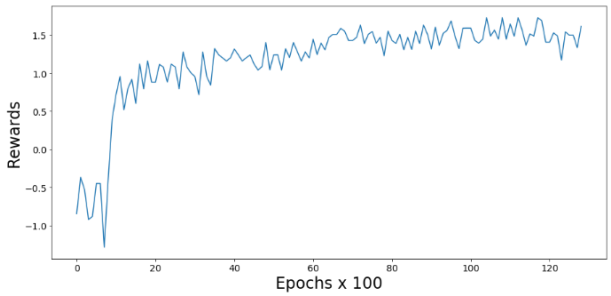

Fig. 5. Rewards vs Epochs for 200k EVs computed over a single run of 13,000 epochs and averaged over 100 steps

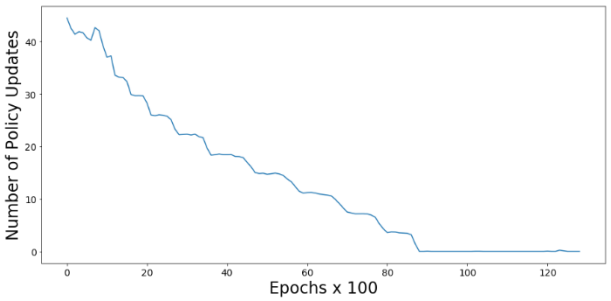

Fig. 6. Policy Updates vs Epochs for 200k EVs computed over a single run of 13,000 epochs and averaged over 100 steps

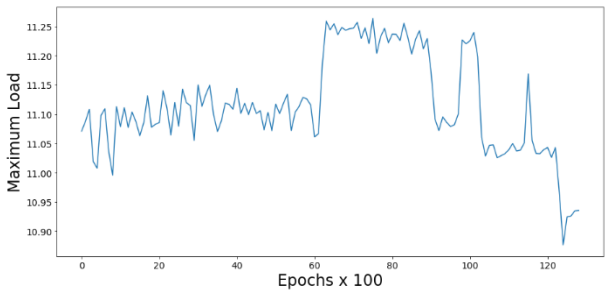

Fig. 7. Maximum Load vs Epochs for 200k EVs computed over a single run of 13,000 epochs and averaged over 100 steps

The PAR appears to decrease, the reward to increase, and the number of updates drops to zero, all indicating convergence. This needs to be run longer to ensure convergence. However, running the simulation further revealed divergence of the policy. This can be seen from the change in average reward, PAR values, and number of updates of the policy for 500k EVs after 1 million epochs, as illustrated in Fig. 8-11.

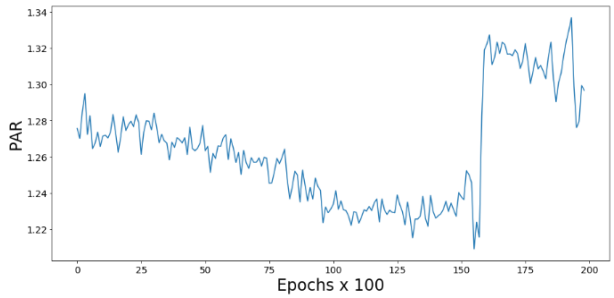

Fig. 8. PAR vs Epochs for 500k EVs computed over a single run of 20,000 epochs and averaged over 100 steps

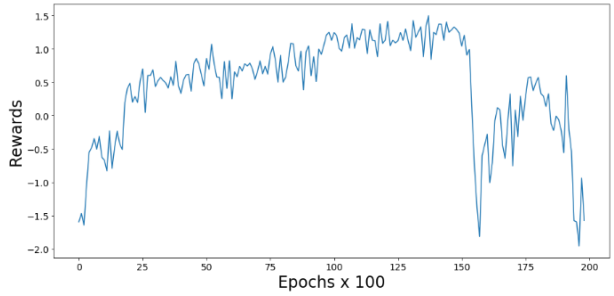

Fig. 9. Rewards vs Epochs for 500k EVs computed over a single run of 20,000 epochs and averaged over 100 steps

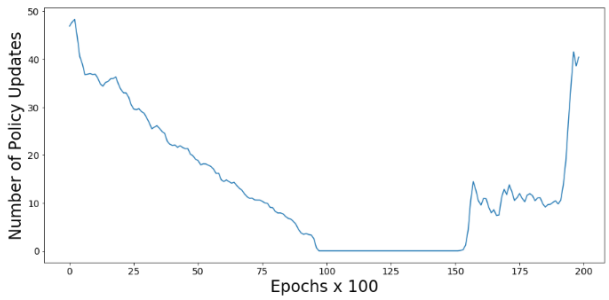

Fig. 10. Number of Updates vs Epochs for 500k EVs computed over a single run of 20,000 epochs and averaged over 100 steps

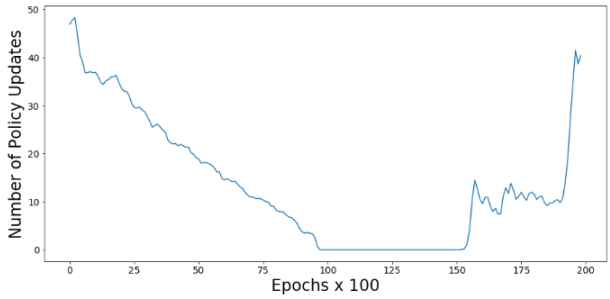

Fig. 11. Maximum Load vs Epochs for 500k EVs computed over a single run of 20,000 epochs and averaged over 100 steps

Running the simulation for 1 million epochs revealed that, multiple times, some time after converging, the reward, and PAR diverged. The PAR started to increase and reward to decrease. The number of updates started to increase. This cycle repeated many times during the 1 million epoch run. To mitigate this we could try to use different temporal difference algorithm that performs better on non-stationary environments, such as Repeated Update Q-Learning [29] or double Q-Learning [23]. To ensure the lack of convergence persists, a simulation for 100k EVs, 400k EVs, and 800k EVs, was performed for 100,000 epochs for 30 runs each. The results are shown in Fig. 12-16. It appears that for 100k EVs the PAR appears to decrease to 1.068, rewards to increase, and the number of policy updates to stabilize. For 400k EVs PAR fluctuated in an un-stable fashion between 1.250 and 1.258 and the rewards were relatevely stable, barely above 0, indicating the policy found could keep the EVs changed up to $S_{V2Gendmin}$ by h_{V2Gend}. For 800k EVs PAR appears stable around 1.510, however the rewards remained stable around -0.5 indicating that the policy could not maintain on average the mean EV SOCs to be $S_{V2Gendmin}$ by h_{V2Gend}. The number of policy updates stabilized around 10 for all 3 cases. The short range in which PAR fluctuated, and the relative convergence of rewards and number of policy updates indicates convergence, however it appears for 800k EV

case the algorithm, on average. is unable to find a satisfactory policy.

To find the best policy a manual inspection was required. The policy that showed good qualitative convergence in a local region was picked for each EV penetration case.

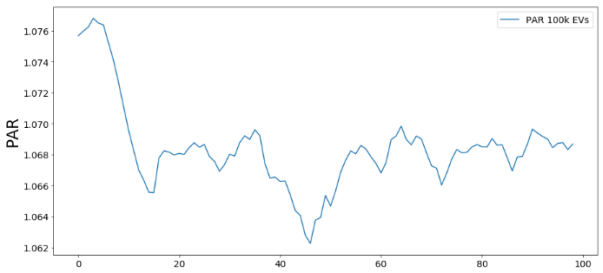

Fig. 12. PAR vs Epochs for 100k EVs computed over 30 runs of 100,000 epochs and averaged over 1000 steps

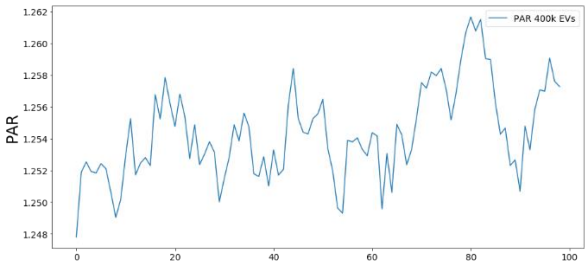

Fig. 13. PAR vs Epochs for 400k EVs computed over 30 runs of 100,000 epochs and averaged over 1000 steps

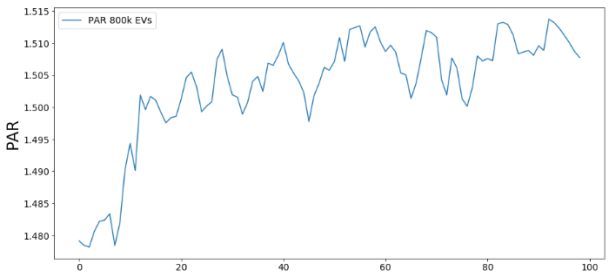

Fig. 14. PAR vs Epochs for 800k EVs computed over 30 runs of 100,000 epochs and averaged over 1000 steps

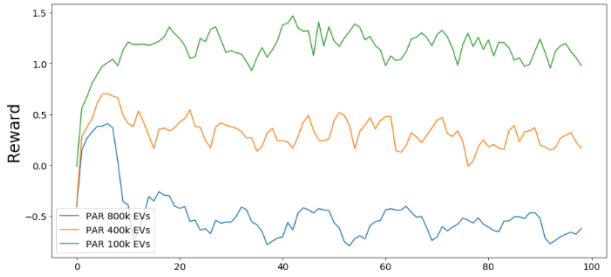

Fig. 15. Reward vs Epochs for 100k, 400k, and 800k EVs computed over 30 runs of 100,000 epochs and averaged over 1000 steps

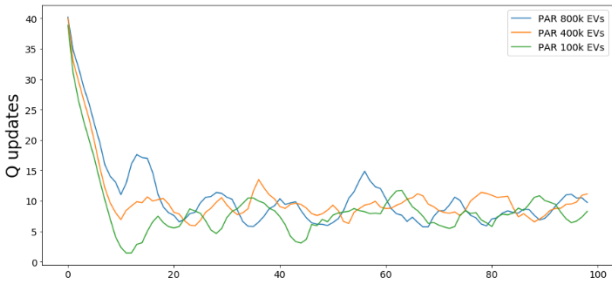

Fig. 16. Number of Updates vs Epochs for 100k, 400k, and 800k EVs computed over 30 runs of 100,000 epochs and averaged over 1000 steps

As can be seen in Fig. 17, after manually selecting action-value functions with converged reward, PAR, and number of policy updates, with, and without V2G service, over different numbers of EVs, V2G scheme does not reduce PAR except in 100k and 200k EV cases.

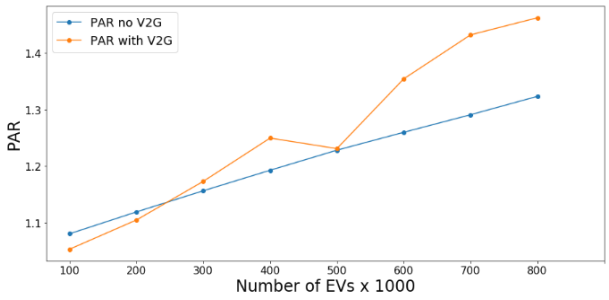

Fig. 17. PAR with V2G and without averaged over 500 trials

The graph of PAR versus the number of EVs using V2G service and without it, averaged over 500 trials, reveals that in all cases except 100k and 200k EV, the PAR resulting from V2G service is not reduced. However, in the 500k EV case the PAR is very close to that without V2G service. It appears the PAR without the V2G service increases linearly with increasing number of EVs.

The load demand levels for cases with V2G scheme and no V2G scheme for 100k EVs and 800k EVs, averaged over 500 trials, are shown in Fig. 18, 19.

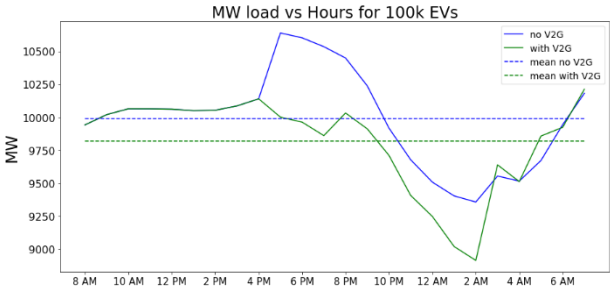

Fig. 18. Load Levels with and without V2G for 100k EVs averaged over 500 trials

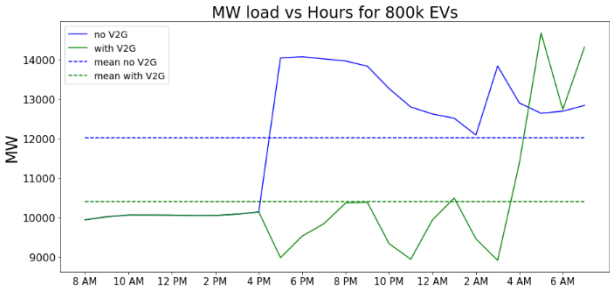

Fig. 19. Load Levels with and without V2G for 800k EVs averaged over 500 trials

It appears V2G scheme removes the peak at 5 PM, but may create another peak at 5 AM. In 100k case, V2G reduces average standard deviation between 8 AM – 7 AM from 363.0 MW down to 344.9 MW, and the standard deviation for the full day from 357.7 MW down to 347.4 MW. In 800k case, V2G reduces average standard deviation between 8 AM – 7 AM from 1641.2 MW down to 1231.3 MW, and the standard devation for the full day from 1615.6 MW down to 1455.1 MW. Average mean SOC of EVs is above 48% with and without V2G. The graph of the total load using V2G and not using V2G reveals that for all EV cases studied the afternoon charging peak present without V2G service is eliminated. This however may create a new peak between 6 AM and 7 AM. The standard deviation of the load is reduced in both 8 AM-until-7 AM and full-day cases in the 800k EV case. Comparing the PAR graph with the load graph it seems the cause of the increasing PAR is the spike due to charging resulting between 5 AM to 7 AM. This is a consequence of the structure of the penalty P_h to ensure adequate SOC at 8 AM. The penalty favors charging in latter hours. Adjusting the penalty may alleviate the issue.

The standard deviation of daily load demand for cases with V2G scheme and no V2G scheme are shown in Fig. 20-23. It appears V2G service generally reduces the demand load standard deviation. From the graph of the standard deviation of the load with V2G service against standard deviation of the load without V2G service, it can be seen that V2G service generally reduces the standard deviation of the demand load.

As can be seen in Fig. 20-23, the standard deviation of the demand load is reduced on average by 2.7% in the case of 24 hours and is reduced on average by 16.2% in the case of the first 23 hours. The standard deviation of demand loads is lower without V2G service for 200k and 300k EVs, but the difference in standard deviation is reduced for 200k and 300k EV cases from about 15% difference to just under 10% difference when counting the first 23 hours of the day.

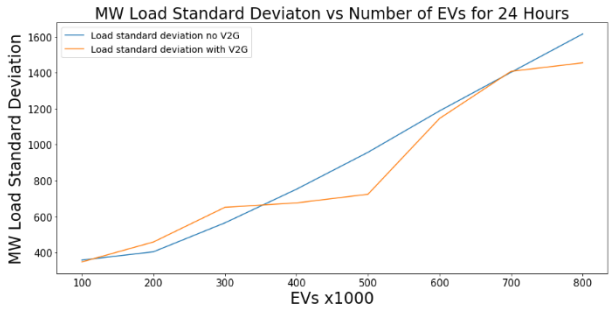

Fig. 20. Load standard deviation with V2G and without for 24 hours averaged over 500 trials

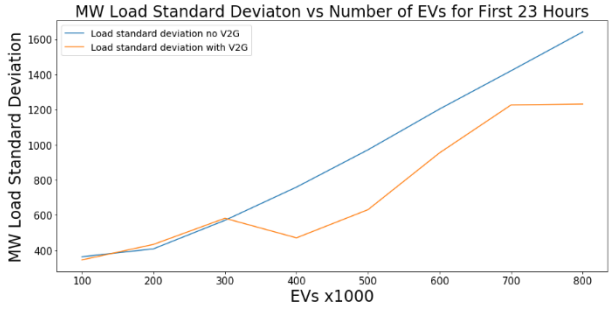

Fig. 21. Load standard deviation with V2G and without for first 23 hours averaged over 500 trials

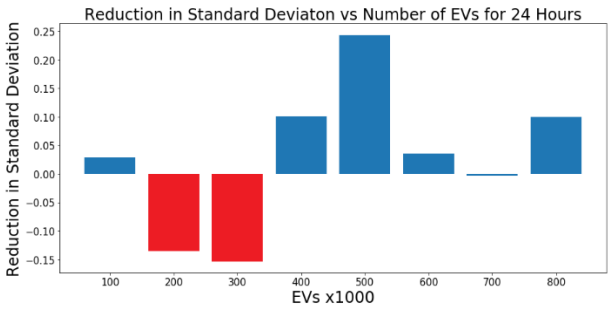

Fig. 22. Load standard deviation reduction with V2G and without for 24 hours averaged over 500 trials

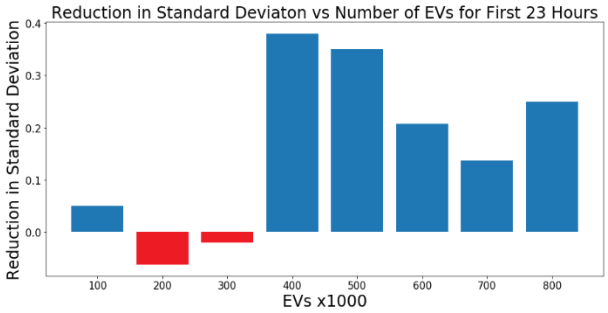

Fig. 23. Load load standard deviation reduction with V2G and without for 23 hours averaged over 500 trials

Optimum policies for participation levels 50% and 75% for 500k EVs are shown in Fig. 24, 25. These policies are not unique, but combinations of charging/discharging are unique. Consecutive discharging actions are not desirable in 0-25% SOC range because they force EVs with low SOC values to discharge but are present in both policies. Different-colored regions depict different actions. There is a strong push to charge between 5 AM and 7 AM.

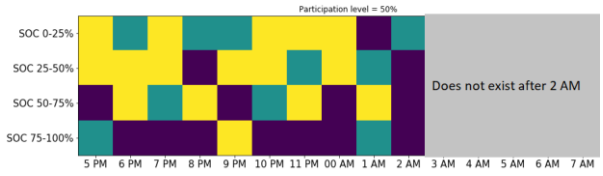

Fig. 24. Optimal policy for 500k EVs with participation level of 50%

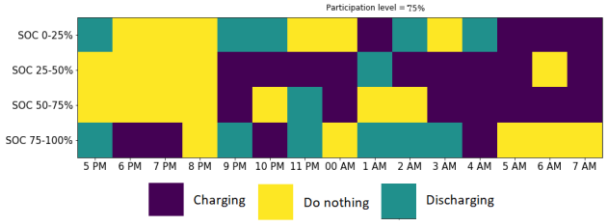

Fig. 25. Optimal policy for 500k EVs with participation level of 75%

The advantages of the proposed method are that EVs can act autonomously knowing only the participation level. Also the load standard deviation for the full day was generally reduced. The disadvantages are that the method only works on constant load demand data, the SOC of EVs at 8 AM is a mean value and on any individual day the mean may be less, it is unclear whether the resulting policy is optimal, and some discharging in consecutive hours is present for EVs with less than 25% charge.

VII. Conclusion

We presented a Q-learning-based algorithm for scheduling EV demand response in the context of a cooperative multi-agent multi-objective game. The method did not reduce PAR but resulted in the average standard deviation reduction of demand load of 2.7% for the full day and 16.2% reduction when the first 23 hours of the day are considered. The convergence of the algorithm after 100,000 epochs of training averaged over 30 runs appears to be clear. The advantages of the method are the reduction in the standard deviation of the load demand, and the ability of each EV to act independently knowing only the participation level for the current time slot. The disadvantages are the possibility of discharging each EV when its state of charge is close to its minimum, and the lack of guarantee that each EV will have a minimum required charge in the morning.

VIII. Future Work

Future work can include changing the underlying algorithm from Q-Learning to Deep Q-Learning utilizing a neural network and generalizing states. The reward definition can be reworked to enable a more balanced charging scheme by placing non-zero rewards for hours close to when the V2G service starts and closer to midnight. Reducing the effect of non-stationarity also can be investigated to mitigate the bias in the environment.

References

[1] "Global EV Outlook 2020 – Analysis - IEA." https://www.iea.org/reports/global-ev-outlook-2020 (accessed Dec. 04, 2020).

[2] "Global electric vehicle stock in the Sustainable Development Scenario, 2019 and 2030 – Charts – Data & Statistics - IEA." https://www.iea.org/data-and-statistics/charts/global-electric-vehicle-stock-in-the-sustainable-development-scenario-2019-and-2030 (accessed Dec. 05, 2020).

[3] G. A. Putrus, P. Suwanapingkarl, D. Johnston, E. C. Bentley, and M. Narayana, "Impact of electric vehicles on power distribution networks," *5th IEEE Veh. Power Propuls. Conf. VPPC '09*, pp. 827–831, 2009, doi: 10.1109/VPPC.2009.5289760.

[4] M. Muratori, "Impact of uncoordinated plug-in electric vehicle charging on residential power demand," *Nat. Energy*, vol. 3, no. 3, pp. 193–201, 2018, doi: 10.1038/s41560-017-0074-z.

[5] "Vehicle-to-grid potential and variable renewable capacity relative to total capacity generation requirements in the Sustainable Development Scenario, 2030 – Charts – Data & Statistics - IEA." https://www.iea.org/data-and-statistics/charts/vehicle-to-grid-potential-and-variable-renewable-capacity-relative-to-total-capacity-generation-requirements-in-the-sustainable-development-scenario-2030 (accessed Dec. 05, 2020).

[6] S. Deb, K. Kalita, and P. Mahanta, "Review of impact of electric vehicle charging station on the power grid," *Proc. 2017 IEEE Int. Conf. Technol. Adv. Power Energy Explor. Energy Solut. an Intell. Power Grid, TAP Energy 2017*, vol. 1, pp. 1–6, 2018, doi: 10.1109/TAPENERGY.2017.8397215.

[7] H. Shareef, M. M. Islam, and A. Mohamed, "A review of the stage-of-the-art charging technologies, placement methodologies, and impacts of electric vehicles," *Renew. Sustain. Energy Rev.*, vol. 64, no. December 2017, pp. 403–420, 2016, doi: 10.1016/j.rser.2016.06.033.

[8] H. K. Nguyen, J. Bin Song, and Z. Han, "Demand side management to reduce Peak-to-Average Ratio using game theory in smart grid," 2012, doi: 10.1109/INFCOMW.2012.6193526.

[9] S. Jaiswal and M. S. Ballal, "Optimal load management of plug-in electric vehicles with demand side management in vehicle to grid application," *2017 IEEE Transp. Electrif. Conf. ITEC-India 2017*, vol. 2018-Janua, pp. 1–5, 2018, doi: 10.1109/ITEC-India.2017.8356942.

[10] H. Liang, B. J. Choi, W. Zhuang, and X. Shen, "Optimizing the energy delivery via V2G systems based on stochastic inventory theory," *IEEE Trans. Smart Grid*, vol. 4, no. 4, pp. 2230–2243, 2013, doi: 10.1109/TSG.2013.2272894.

[11] D. S. Lokunarangoda and I. A. Premaratne, "Optimum charging algorithm for electric vehicles to reduce the peak demand in consumer premise," *2017 Int. Conf. Comput. Commun. Electron. COMPTELIX 2017*, pp. 654–658, 2017, doi: 10.1109/COMPTELIX.2017.8004050.

[12] K. Mahmud, M. J. Hossain, and G. E. Town, "Peak-Load Reduction by Coordinated Response of Photovoltaics, Battery Storage, and Electric Vehicles," *IEEE Access*, vol. 6, pp. 29353–29365, 2018, doi: 10.1109/ACCESS.2018.2837144.

[13] N. Erdogan, F. Erden, and M. Kisacikoglu, "A fast and efficient coordinated vehicle-to-grid discharging control scheme for peak shaving in power distribution system," *J. Mod. Power Syst. Clean Energy*, vol. 6, no. 3, pp. 555–566, 2018, doi: 10.1007/s40565-017-0375-z.

[14] G. Zhang, S. T. Tan, and G. Gary Wang, "Real-Time Smart Charging of Electric Vehicles for Demand Charge Reduction at Non-Residential Sites," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4027–4037, 2018, doi: 10.1109/TSG.2016.2647620.

[15] J. Singh and R. Tiwari, "Multi-Objective Optimal Scheduling of Electric Vehicles in Distribution System," *2018 20th Natl. Power Syst. Conf. NPSC 2018*, 2018, doi: 10.1109/NPSC.2018.8771768.

[16] K. Mahmud, M. J. Hossain, and J. Ravishankar, "Peak-Load Management in Commercial Systems with Electric Vehicles," *IEEE Syst. J.*, vol. 13, no. 2, pp. 1872–1882, 2019, doi: 10.1109/JSYST.2018.2850887.

[17] S. Mocci, N. Natale, F. Pilo, and S. Ruggeri, "Multi-agent control system for the exploitation of Vehicle to grid in active LV networks," 2016, doi: 10.1049/cp.2016.0813.

[18] R. Xiong, J. Cao, and Q. Yu, "Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle," *Appl. Energy*, vol. 211, no. 5, pp. 538–548, 2018, doi: 10.1016/j.apenergy.2017.11.072.

[19] S. Dimitrov and R. Lguensat, "Reinforcement Learning Based Algorithm for the Maximization of EV Charging Station Revenue," *Proc. - 2014 Int. Conf. Math. Comput. Sci. Ind. MCSI 2014*, pp. 235–239, 2014, doi: 10.1109/MCSI.2014.54.

[20] Y. Cao, D. Li, Y. Zhang, and X. Chen, "Joint Optimization of Delay-Tolerant Autonomous Electric Vehicles Charge Scheduling and Station Battery Degradation," *IEEE Internet Things J.*, 2020, doi: 10.1109/JIOT.2020.2992133.

[21] X. Fang, J. Wang, G. Song, Y. Han, Q. Zhao, and Z. Cao, "Multi-agent reinforcement learning approach for residential microgrid energy scheduling," *Energies*, 2019, doi: 10.3390/en13010123.

[22] "Alberta Electric System Operator." http://ets.aeso.ca/ (accessed Oct. 15, 2020).

[23] A. G. Sutton, R. S., Barto, *Reinforcement Learning: An Introduction*. The MIT Press., 2018.

[24] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *arXiv*. 2019.

[25] D. Lee, N. He, P. Kamalaruban, and V. Cevher, "Optimization for Reinforcement Learning: From a single agent to cooperative agents," *IEEE Signal Process. Mag.*, 2020, doi: 10.1109/MSP.2020.2976000.

[26] C. Liu, X. Xu, and D. Hu, "Multiobjective reinforcement learning: A comprehensive overview," *IEEE Trans. Syst. Man, Cybern. Syst.*, 2015, doi: 10.1109/TSMC.2014.2358639.

[27] Y. Gong, M. Abdel-Aty, J. Yuan, and Q. Cai, "Multi-Objective reinforcement learning approach for improving safety at intersections with adaptive traffic signal control," *Accid. Anal. Prev.*, 2020, doi: 10.1016/j.aap.2020.105655.

[28] S. Übermasser and M. Stifter, "A multi-agent based approach for simulating G2V and V2G charging strategies for large electric vehicle fleets," 2013, doi: 10.1049/cp.2013.0959.

[29] S. Abdallah and M. Kaisers, "Addressing environment non-stationarity by repeating Q-learning updates," *J. Mach. Learn. Res.*, 2016.