Grid Stability Management and Price Arbitrage for Distributed Energy Storage and Generation via Reinforcement Learning

Qizhan Tam, *Stanford University*

***Abstract*—A major concern with distributed energy storage and generation is the resulting grid instability. Making this situation worse is the possibility of arbitrage using batteries. We developed a method using Q-learning to balance the benefits of arbitrage with the cost of voltage violation. We find that we can successfully find a battery operation policy that balances both components through weighting their respective reward functions.**

I. Introduction

As distributed energy storage and generation becomes more widespread, there is a growing concern that the variability of power supply and demand from such sources will adversely affect grid stability. Previous work on using reinforcement learning for distributed storage and generation had been limited to either arbitrage using batteries without considering the adverse effects on the grid[1], or a personal home energy management system[2]. A major downside of managing distributed resources without considering their impact on the grid is that it causes voltage violation, that is when the grid voltage exceeds +/- 5% of the operational voltage. Voltage violation has severe effects on grid and household electrical equipment as they are usually only rated to operate within the +/- 5% operational voltage window. For this project, we present a method that uses Q-learning to find a battery operation policy that balances voltage violation cost with the arbitrage returns.

II. Grid Simulation Model

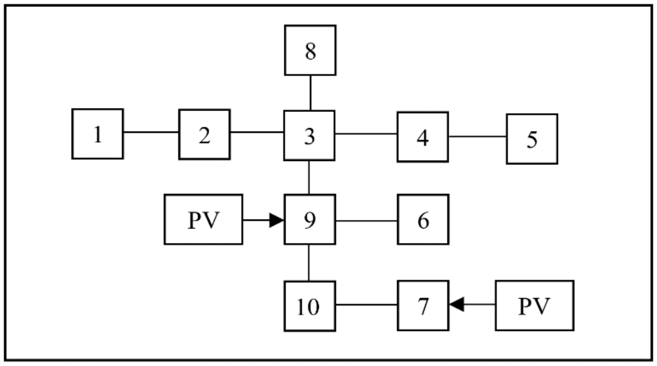

Fig. 1. The distribution grid modelled as node network with Node 1 being the substation.

A. Distribution Test Feeder

This study utilizes the IEEE 123-bus feeder (IEEE-123) developed by the IEEE PES Distribution System Analysis Subcommittee's Distribution Test Feeder Working Group. As its name implies, IEEE-123 is a radial node test feeder model with 123 bus or also referred to as "nodes". This paper will use the latter terminology. The node structure is seen in Figure 1. For this study, we simplified IEEE-123 to a reduced order model by using the first 10 nodes and their respective branches (Figure 2). The first node represents the substation, where high voltage lines are stepped down to the nominal voltage of 4.16 kV. The branches are modelled as overhead power lines with distances between each node taken into account when calculating resistances. At each node, voltage is again stepped down to ~120V by transformers before fed into the laterals that supply power to houses.

B. Power Flow Simulation

MATPOWER, a package consisting of MATLAB M-files, is used to simulate the grid model's power flow. As the residential load data (see part C) is in terms of real power, a power factor in the range of 0.875 to 0.925 is randomly generated for each node at each time-step to account for the reactive power component of the load.

C. Residential Load

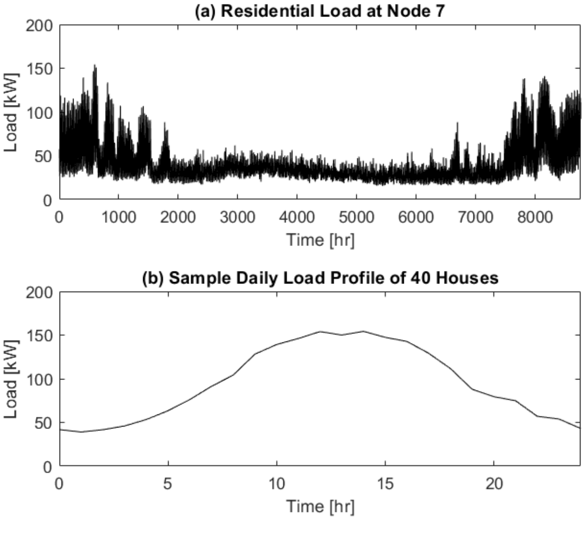

Fig. 2. (a) The low load period corresponds to fall and winter. The high load period corresponds to spring and summer. (b) The maximum load usually occurs during midday.

A key component of this study that differs from previous work that used reinforcement learning in grid storage problems is that we take into account the variable nature of individual houses' power demand. Each node was randomly assigned the load of 40 houses from a dataset of 2000 houses in Bakersfield from 8/1/2010 to 7/31/2011.

III. Price Arbitrage Problem

Using Q-learning on price arbitrage for battery storage have been explored previously by Wang et al. (2017)[1]. The method used in the paper was modified for the purposes of this study. The arbitrage optimization problem is formulated as below:

$$\underset{a_B}{argmax} \sum_{t=0}^{T} p_t\, c_B\, a_{B,t}$$

The goal is to find the optimal operating policy for a battery with power capacity of c_B based on the current electricity price, $p(t)$. The policy will determine the battery operation, a_B, that will maximize the cumulative return by charging the battery during low electricity prices and discharging the battery at high electricity prices.

A. State Space

The state space of this problem is described by the current electricity price, $p(t)$ [\$/MW], and the battery state, s_B [MW]. The electricity price used is the real-time hourly Locational Marginal Price (LMP) from 8/1/2016 to 7/31/2017 of the PHILADRD price node provided by PJM. LMP accounts for the system energy price, transmission congestion cost, marginal losses and effect of reserve shortages. As such, LMP is commonly used to reflect the value of electric energy at different places. Electricity price is taken as a spot price, where the price for the next hour is known at the start of the hour.

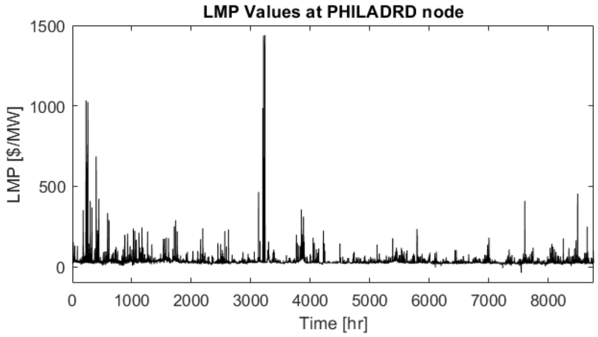

Fig. 3. In contrast with the residential loads, LMP values don't have any noticeable temporal patterns.

The state space is discretized by dividing the LMP values into 1000 equally-sized intervals. A battery with a capacity $c_B = 1$ [MW] is placed at node 7 of Figure 1. The battery has a charging rate of 1 C, which means it takes an hour for the battery to fully charge. As such, the battery only has 2 discrete states:

$$s_B = \begin{cases} 1\text{ MW} & \text{if Fully Charged} \\ 0\text{ MW} & \text{if Fully Discharged} \end{cases}$$

B. Action Space

The only actions, a_B, available are related to the battery, with the set of actions being:

$$a_B = \begin{cases} 1 & \text{if Charge} \\ 0 & \text{if Idle} \\ -1 & \text{if Discharge} \end{cases}$$

The effects of the actions on the battery state is:

$$s_{B,t} = s_{B,t} + a_{B,t}\, c_B$$

C. Rewards

The difference between the average price $\overline{p_t}$ and the current price p_t is used to calculate the reward. This is to avoid conservative actions and to compare the current price to historical values.

$$r_{B,t} = \begin{cases} (\overline{p_t} - p_t)\, c_B & \text{if Charge} \\ 0 & \text{if Idle} \\ (p_t - \overline{p_t})\, c_B & \text{if Discharge} \end{cases}$$

D. Algorithm

Algorithm 1: Q-learning for Electricity Price Arbitrage

```
function ARBITRAGE
t ← 0
$s_{B,0}$ ← 0
Q ← 0
loop
    $s_t \leftarrow [s_{B,t}, p_t]$
    if ϵ < randomly generated number from 0 to 1 then
        $a_t$ ← randomly choose possible action
    else
        $a_t \leftarrow argmax_a\, Q(s_t, a_t)$
    $r_{B,t} \leftarrow (\bar{p}_t - p_t)\, c_B\, a_{B,t}$
    $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_t + max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$
    t ← t + 1
until t = T
return Q
```

IV. Grid Stability Problem

A 3 MW PV plant is placed at node 7 and a 2 MW PV plant is placed at node 9 as depicted in Figure 1. This problem is to minimize voltage violations for a given load, P_t in the grid model. As a negative cost function is used to evaluate voltage violation, vv_t, at each time-step, the problem is formulated as below:

$$\underset{a_B}{argmax} \sum_{t=0}^{T} r_t(vv_t\,(P_t, a_{B,t})\,)$$

A. State Space

The state space of this problem consists of the combined time-averaged residential load of node 7 and the PV plant's power output for the next hour, and the battery's state of charge.

To model the PV plant's power output, solar radiation data of Bakersfield from 8/1/2010 to 7/31/2011 was obtained from

NREL's National Solar Radiation Database. This set of data was generated from the Physical Solar Model (PSM), which accounts for cloud properties from satellite data. From the solar radiation data, irradiance, $E_t\ [W/m^2]$, is used to calculate the power output, $PV_t\ [W]$:

$$PV_t = \eta * E_t * A * N_{panels}$$

Astronergy's polycrystalline CHSM6610P Series 275 W PV module is used as a model to find the parameters for the equation above: η = 0.169, A = 1.408 m^2, N_{panels} = 12501 for node 7 and N_{panels} = 8334 for node 9. The battery's state of charge remains the same as the arbitrage problem.

B. Action Space

The action space of this problem is the same as the price arbitrage problem.

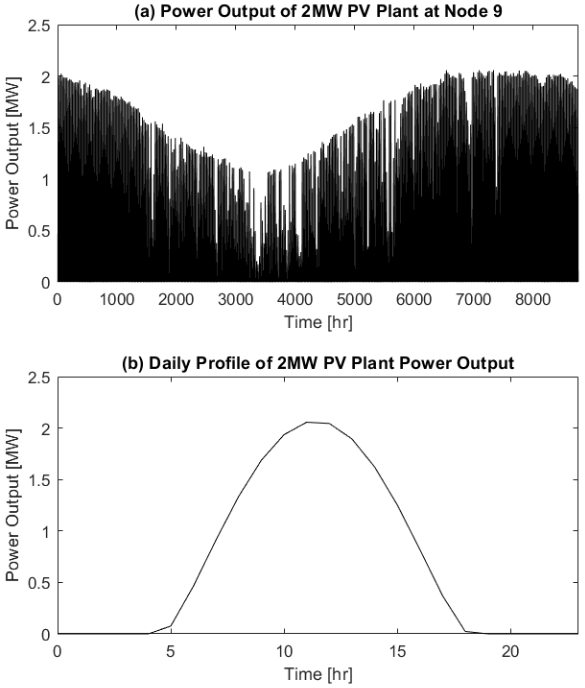

Fig. 4. (a,b) The power output follows seasonal and daily patterns with low power output corresponding to times of low solar radiation and vice versa.

C. Rewards

Voltage violation occurs when the voltage of the system, v_t exceeds $+/-$ 5% of the operational voltage. The operational voltage in this study is normalized to 1 and so the level of voltage violation is calculated:

$$vv_t = \begin{cases} 0 & \text{if } |v_t| < 1 \\ |v_t - 1| & \text{if } |v_t| > 1 \end{cases}$$

As the adverse consequences to the grid components increase with greater voltage violation, a linearly increasing cost function is used:

$$r_{vv,t} = -1000 * vv_t - 50$$

D. Algorithm

Algorithm 2: Q-learning for Grid Stability

```
function PowerFlow ($P_t, a_t$)
$v_t$ ← MatPOWER ($P_t, a_t$)
if $|v_t|$ < 1
    $vv_t$ ← 0
else
    $vv_t$ ← $|v_t - 1|$
return $vv_t$
function GridStability
t ← 0
$s_{B,0}$ ← 0
Q ← 0
loop
    $s_t$ ← $[s_{B,t}, P_t]$
    if ϵ < randomly generated number from 0 to 1 then
        $a_t$ ← randomly choose possible action
    else
        $a_t$ ← $argmax_a\ Q(s_t, a_t)$
    $vv_t$ ← PowerFlow ($P_t, a_t$)
    $r_{vv,t}$ ← $-1000 * vv_t - 50$
    $Q(s_t, a_t)$ ← $Q(s_t, a_t) + \alpha(r_{vv,t} + max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$
    t ← t + 1
until t = T
return Q
```

V. Price Arbitrage and Grid Stability Problem

A. State Space

The state space is a combination of both problems: $s_t = [s_{B,t}, p_t, P_t]$. This means the state space has dimensions of 2x100x1000.

B. Action Space

The action space remains the same as the previous two problems.

C. Reward

The reward is the sum of the two rewards:

$$r_t = r_{B,t} + r_{vv,t}$$

The relative weights of the two rewards can be adjusted to give more importance to arbitrage or preventing voltage violation. As voltage violation can cause severe damage to grid and household electrical equipment, we penalize voltage violation much greater than the potential return from arbitrage.

D. Algorithm

The algorithm for the combined problem is similar to Algorithm 2 with modifications to the reward function.

Algorithm 3: Q-learning for Arbitrage & Grid Stability

```
function COMBINEDPROBLEM
t ← 0
$s_{B,0} \leftarrow 0$
Q ← 0
loop
    $s_t \leftarrow [s_{B,t}, p_t, P_t]$
    if ϵ < randomly generated number from 0 to 1 then
        $a_t \leftarrow$ randomly choose possible action
    else
        $a_t \leftarrow argmax_a\, Q(s_t, a_t)$
    $vv_t \leftarrow$ POWERFLOW $(P_t, a_t)$
    $r_{vv,t} \leftarrow -1000 * vv_t - 50$
    $r_{B,t} \leftarrow (\bar{p}_t - p_t)\, c_B\, a_{B,t}$
    $r_t \leftarrow r_{vv,t} + r_{B,t}$
    $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_t + max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$
    t ← t + 1
until t = T
return Q
```

VI. Roll-Out

After each Q-learning algorithm is run for 3000 iterations, the Q-values generated are then used to determine the battery operation policy using Algorithm 4.

Algorithm 4: Calculating the returns of a policy.

```
function ROLLOUT
t ← 0
$s_{B,0} \leftarrow 0$
$Q \leftarrow$ ARBITRAGE/GRIDSTABILITY/COMBINEDPROBLEM
loop
    $s_t \leftarrow [s_{B,t}, p_t, P_t]$
    $a_t \leftarrow argmax_a\, Q(s_t, a_t)$
    $vv_t \leftarrow$ POWERFLOW $(P_t, a_t)$
    $r_{vv,t} \leftarrow -1000 * vv_t - 50$
    $r_{B,t} \leftarrow (\bar{p}_t - p_t)\, c_B\, a_{B,t}$
    $r_t \leftarrow r_{vv,t} + r_{B,t}$
    t ← t + 1
until t = T
return $r_{vv,t}, r_{B,t}, r_t$
```

VII. Results & Discussions

Q-learning Algorithm	Cost of Voltage Violation	Incidents of Voltage Violation	Arbitrage Return
none	$ -17187	1699	$ 0
1	$ -19977	1969	$ 11478
2	$ -15656	1547	$ 5004
3	$ -16081	1586	$ 23598

Table 1. Results from ROLLOUT using Q-values from the different Q-learning algorithms and a control run where there is no battery operation.

A. *Grid Stability (Voltage Violation)*

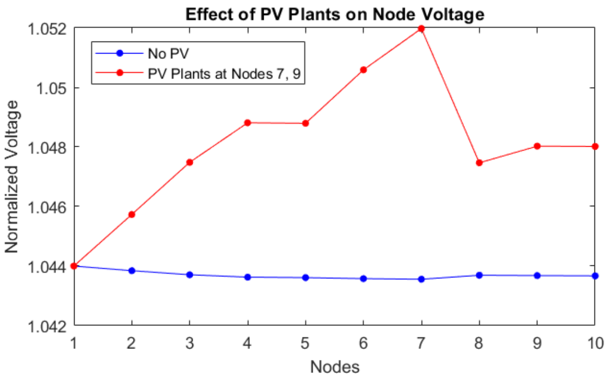

Fig. 5. (a,b) The power output follows seasonal and daily patterns with low power output corresponding to times of low solar radiation and vice versa.

Power output from the PV plants increases the overall voltage of the grid, with the nodes closest to the PV plants being affected the most. The first node (substation) is the only one not affected as voltage regulation at substations. The further away from a node is from Node 1, the less optimal its voltage will be. Normally, in large networks with hundreds of nodes, voltage drop is the main cause of voltage violation. However, as demonstrated in the above plot, even for a small grid network, distributed generation could cause voltage violation in the form on overvoltage.

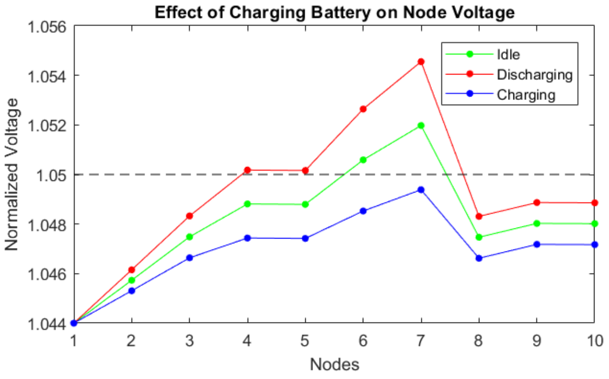

Fig. 6. A charging battery helps reduce voltage violation and a discharging battery increases overvoltage.

The reduction/increase in voltage from a charging/discharging battery is not isolated to the battery's node, the effects are also felt by the other nodes. The further away a node is, the less the battery's effect has on the node's voltage.

If the policy is optimized by only considering the returns from LMP (Algorithm 1), there is a 16% increase in voltage violation incidents compared to the control. By focusing only on the cost of voltage violation, Algorithm 2 provides the best battery operation policy, with a 9% reduction in voltage violation incidents. As expected, the algorithm that considers the benefit of arbitrage and cost of voltage violation recorded the second best reduction in voltage violation incidents at 7% compared to the control case.

B. *Arbitrage Return (LMP)*

The policy derived from Algorithm 2 gave the least return as it only considers the cost of voltage violation. The arbitrage return increases by 229% when Algorithm 1 is used as it only considers rewards derived from LMP. Interestingly, the algorithm that considers both the benefit of arbitrage and cost of voltage violation gave the best return at an increase of 206% compared to Algorithm 1's return. There are two possible reasons for this result: i) the hyper-parameters used for Q-learning are sub-optimal, ii) the "average-price" method recommended by Wang et al. to encourage exploration is not effective in this context.

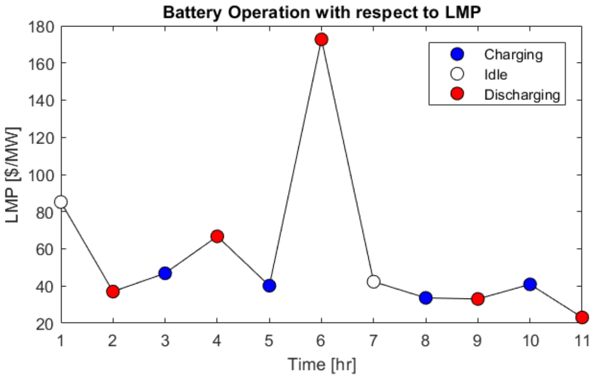

Fig. 7. The battery mostly discharges at high prices and charges at low prices.

VIII. Conclusion

We have demonstrated a method to approximate a policy for battery operation that optimizes for both the arbitrage return and cost of voltage violation. For future work, there is still more work that needs to be done on finding a more optimal way of optimizing for arbitrage returns as it underperforms compared to the combined problem. We can also look into placing batteries at multiple nodes to further mitigate voltage violation and increase arbitrage returns

Acknowledgment

I would like to thank the teaching team of AA 228 for sharing their enthusiasm and guiding me through the course of this project. I would also like to thank the members of the Stanford GISmo Lab especially Dr. Raffi Sevlian and Dr. Michaelangelo Tabone for the fruitful discussions we had regarding this project.

References

1. Wang, Hao et al. "Energy Storage Arbitrage in Real-Time Markets Via Reinforcement Learning." Preprint, 2017, arXiv:1711.03127.pdf
2. Mocanu, Elena et al. "On-line Building Energy Optimization using Deep Reinforcement Learning." Preprint, 2017, arXiv:1707.05878.

* Files used for this project can be downloaded from:
https://stanford.box.com/s/f4iyu4tjnxc6ald5h63g791czeb8o5ue