NAME: Micah Botkin-Levy
DATE: 4/18/19
Final Project Article
STAT/CS 387

1 Introduction

Energy storage is becoming an important aspect of the electrical grid as prices of lithium-ion batteries and other technologies begin to decrease. With the increased adoption of renewable energy, there is an increase in the demand for flexible resources that can temporarily "move" the electricity by storing or controlling them. Since most renewable energy, such as solar and wind, are nondispatchable and intermittent, energy storage is one solution that can fill these holes in electricity production when the wind stops blowing or the sun stops shining. One method for tracking how available electricity production is compared with the needed electricity demand is having pricing signals. Electric pricing signals can come in the form of real-time prices at the wholesale level or time of use pricing at the consumer level which blocks out times where electricity is cheap or expensive.

Energy storage has many possible services it can provide such as backup power but a major revenue source is using energy storage for real-time temporal arbitrage [1, 2]. This basically means that the energy storage device will charge at low energy prices and discharge at high energy prices. While this is easy to do when the prices are known ahead of time, deciding when to charge and discharge becomes more difficult when there is some uncertainty such as with real-time electricity prices.

Reinforcement learning is often a popular and apt approach for the energy storage control problem as the algorithm needs no forecast of future energy prices. Q-learning, a reinforcement learning algorithm, is particularly appropriate as it is a model-free approach and tells an agent what action to take based on potential rewards. The theory of reinforcement learning and in particular, the Q-learning method, has been around for awhile. Christopher Walkins and Richard Sutton both works on its development in the 1980s and 1990s [3, 4, 5, 6, 7].

Q-learning methods have been used to solve the energy storage system arbitrage problem in previous papers such as *Energy Storage Arbitrage in Real-Time Markets via Reinforcement Learning* [8]. The authors implement a Q-learning algorithm and use a ϵ-greedy method to solve an online version of an energy storage arbitrage maximization problem. Similarly, in *Reinforcement Learning-Based Control of Residential Energy Storage Systems for Electric Bill Minimization* [9], a TD(λ)-learning algorithm is used to control energy storage and compares it to an on-peak, off-peak baseline strategy. The paper [10] also uses Q learning to solve a energy system control problem but expands the algorithm to have a 2 time step look-ahead. Other papers have tackled energy storage control problem using deep Q-learning [11, 12, 13] which expands the algorithm to use a neural network [14]. Q-learning has been applied to other control problems in the energy industry such as smart energy buildings [15] and microgrids [12, 13, 16, 17].

In addition, other researchers have approached the energy storage system control problem

using different methods with the goal to maximize the arbitrage profits. When formed into an optimization problem, a optimization solver can find the optimal schedule of charge and discharge powers if the prices are known ahead of time [17, 18, 19]. In [20], the authors use a dynamic programming approach and incorporate stochastic electricity prices and fluctuations in energy demand to maximize the value of storage.

The article will contribute to research in the area in a couple of ways. First, it will assess if a reinforcement learning algorithm can be used to effectively control an energy storage device under uncertainty of five minute real-time prices. Then, it will be used to assess the benefits and downfalls of certain algorithms for this implementation. Finally, there will be opportunities to determine what future work could be done in this area.

The goal of this project was to research, design, and implement a strategy for on-line energy storage control based on real-time prices. In order to do this, we pulled historical data for real-time prices from ISO New England. Then, we compared reinforcement learning algorithms to determine the actions of the energy storage device. Finally, we assessed the algorithms' performance compared to an optimized control algorithm and random decision baseline.

In Section 2, we will go over the formulation of the energy storage system problem including the energy dynamics, physical constraints, and cost function. Following this in Section 3, we will explain a perfect foresight optimization approach to solving the problem and a random action approach which are used as benchmarks as well as two variations of a reinforcement learning approach. In Section 4, we present and compare the results from these algorithms for a certain scenario. Finally, in Section 5, we present some conclusions and opportunities for future work.

2 Problem Formulation

The problem in question concerns an utility energy storage system that is connected to the electricity grid and is subject to real time prices. For our purpose, we will consider an energy storage system operating over a finite time horizon $k \in \{0, \cdots, K\}$. The goal is to implement a control system for the energy storage systems in order to charge and discharge to perform energy arbitrage and produce a profit for the owners.

2.1 Energy Storage System Dynamics

We are concerned with modeling the energy state at each time step k, $e[k]$, for the energy storage system. The most simple linear discrete time equation for the energy state of the energy storage system is a function of the energy storage system charging power, $p_c[k]$, and the energy storage system discharging power, $p_d[k]$ which have units of megawatts.

$$e[k+1] = e[k] + (p_c[k] - p_d[k])\Delta t \tag{1}$$

for all time steps $k = 0, \cdots, K-1$, time step length Δt in hours, and initial energy $e[0] = e_0$. In this model, we will assume that p_d and p_c are constant across the whole time step. Based

on the physical characteristics of the energy storage system, there are limits on the charging and discharging power as well as the energy capacity.

$$0 \le p_c[k] \le p_c^{\max} \qquad \forall k = 0, \cdots, K-1 \tag{2}$$
$$0 \le p_d[k] \le p_d^{\max} \qquad \forall k = 0, \cdots, K-1 \tag{3}$$
$$0 \le e[k] \le e^{\max} \qquad \forall k = 0, \cdots, K \tag{4}$$

In addition, there are inefficiencies in the energy storage system that we would like to incorporate into our model. First, energy storage systems can not hold a charge indefinitely and this phenomenon is modeled in a self-discharge term in the energy dynamics equations [21]. The self-discharge is lost as a function of the percentage of the current energy state. A new energy dynamics equation now replaces equation (1).

$$e[k+1] = e[k] + (p_c[k] - p_d[k])\Delta t - \frac{sd}{100} e[k]\Delta t \tag{5}$$

where sd is the percentage of energy that is lost due to self-discharge per hour. Another inefficiency is due to the loss of power in the exchange between the energy grid and the energy storage system. There are inefficiencies that make the power purchased from the grid, $p_p[k]$, more than the power that is used to charge the battery, hence $p_p[k] > p_c[k]$. Likewise, the power that is sold to the grid, $p_s[k]$, will always be less than the power discharged, $p_s[k] < p_d[k]$. The relationships can further be defined as

$$p_p[k] = \frac{1}{\eta_c} p_c[k] \tag{6}$$
$$p_s[k] = \eta_d p_d[k] \tag{7}$$

where $\eta_c \in (0,1)$ and $\eta_d \in (0,1)$ are the efficiencies associated with charging and discharging, respectively.

2.2 Energy Arbitrage Cost Function

The objective of the energy storage control is to profit from the sale and purchase of electricity based on the real time electricity prices. We will define $x[k] \in \mathbb{R}$ as the real time price of electricity at time step k which has units of $\frac{\$}{MWh}$. We can now create an equation for the cost to the energy storage system in terms of the power charged or discharged during this time step.

$$\text{cost}[k] = x[k](p_p[k] - p_s[k])\Delta t \tag{8}$$

Replacing these terms with the charging and discharging variables using equations (6) and (7) gives us

$$\text{cost}[k] = x[k](\frac{1}{\eta_c} p_c[k] - \eta_d p_d[k])\Delta t \tag{9}$$

The solution to this problem will attempt to minimize this cost function in order to maximum profits of the energy storage system.

2.3 Simultaneous Charging and Discharging

Astute readers will notice that we have imposed no constraints on preventing charging and discharging during the same time step. Physically, an energy storage system cannot charge and discharge energy at the same time. Since we have assumed that the charging and discharging power is constant across the time step, this infers that we should include a constraint

$$p_d[k]p_c[k] = 0 \quad \forall k = 0, \cdots, K-1 \tag{10}$$

which imposes that either $p_d[k] = 0$ or $p_c[k] = 0$ or both are zero. However because the efficiencies (η_c and η_d) are both assumed to be less than zero, it will always be optimal to either discharge or charge at each timestep depending on the real time energy price.

3 Problem Approaches

In this section, we will cover different approaches to solving this control problem. We begin by formulating an optimization problem that can be used in the idealize situation of a perfect forecast of energy prices. Then we will look at reinforcement algorithms which are more realistic for an online implementation with no forecast of energy prices.

3.1 Perfect Forecast Optimization Algorithm

In an ideal world, we would have a perfect prediction of the real time energy prices for the entire time horizon. In this case, an optimization problem could find the best schedule of charging and discharging powers in order to maximum profits. To formulate this problem we want to minimize the cost objective function in equation (9) subject to the energy storage system dynamics. Note that this is called a linear problem since it has a linear objective function subject to linear constraints. Although the problem has a lot of variables due to the large horizon, linear problems are easily solved and well behaved.

While solving this optimization problem is useful to use as a best case cost over the time horizon, this approach would not work for implementing as a real time control algorithm. This is because in reality, real time energy prices are very volatile and hard to predict. In order to solve this problem and maximum profits in real time another approach will be needed.

3.2 Reinforcement Learning Algorithm

Reinforcement learning provides a efficient solution to the problems in which (i) actions depend on current system states and affect future states; (ii) return is optimized cumulatively; (iii) Markov Processes; (iv) the system might be non-stationary [7]. These properties make reinforcement learning different from other machine learning techniques and model-based optimization. The energy storage problem described above meets all four properties (i) actions (charge/discharge) depend on current real time energy prices and actions affect the future

Algorithm 1 Perfect Forecast Optimization Algorithm

Require: Initialize e_0

1: Solve optimization problem:

$$\min \sum_{k=0}^{K-1} x[k](\frac{1}{\eta_c}p_c[k] - \eta_d p_d[k])\Delta t \tag{11a}$$

s.t. (11b)

$$e[k+1] = e[k] + (p_c[k] - p_d[k])\Delta t - \frac{sd}{100}e[k]\Delta t \qquad \forall k = 0, \cdots, K-1 \tag{11c}$$

$$0 \le p_c[k] \le p_c^{\max} \qquad \forall k = 0, \cdots, K-1 \tag{11d}$$

$$0 \le p_d[k] \le p_d^{\max} \qquad \forall k = 0, \cdots, K-1 \tag{11e}$$

$$0 \le e[k] \le e^{\max} \qquad \forall k = 0, \cdots, K-1 \tag{11f}$$

$$e[0] = e[K] = e_0 \tag{11g}$$

states (energy level); (ii) energy storage system's goal is minimize cost for all time steps; (iii) energy storage system does not have a priori knowledge of energy prices but knows past history; (iv) price profiles are non stationary. Therefore, reinforcement learning is a suitable approach for the energy storage system arbitrage problem.

3.2.1 State Space

The state of the system can be defined by the current energy state of the energy storage system $e[k]$ and the current real time price $x[k]$. In order to properly manage the states we chose to discretize the price into X intervals and the energy into E intervals.

$$\mathcal{S} = \{1, \cdots, X\} \times \{1, \cdots, E\} \tag{12}$$

where $\{1, \cdots, E\}$ represents E even intervals ranging from 0 to $e^{\max}$ and $\{1, \cdots, X\}$ represents X intervals defined by X even quantiles of a realistic range of prices.

3.2.2 Action Space

The energy storage system has three actions it can take at each time step

$$\mathcal{A} = \{-\tilde{p}_d, 0, \tilde{p}_c\} \tag{13}$$

where $\tilde{p}_d$ and $\tilde{p}_c$ are the maximum allowable discharge and charging powers based on the energy state of the system. Choosing action $\tilde{p}_d$ is either equivalent to discharging at maximum power $p_d^{\max}$ for the entire time step or the equivalent of discharging at maximum power until the energy storage system reaches a zero energy state. Similarly, choosing action $\tilde{p}_c$ is either equivalent to charging at maximum power $p_c^{\max}$ or the equivalent of charging at maximum power until the energy storage system reaches the maximum energy $e^{\max}$.

3.2.3 Reward 1

We are now concerned with developing a reward function to tell the energy storage system how good the choice of action was given its current energy state and energy price. The simplest choice for this function would just be the negative of our cost function from equation (9).

$$r_1[k] = -\text{cost}[k] = x[k](\eta_d \tilde{p}_d[k] - \frac{1}{\eta_c}\tilde{p}_c[k])\Delta t \tag{14}$$

With this reward function, the energy storage system will receive a positive reward if the real time price $x[k]$ is positive and the energy storage system is discharging, i.e $x[k]\eta_d\tilde{p}_d[k]\Delta t$. Likewise, if the energy storage system is charging and the real time price is positive it will receive a negative reward. Conversely, these rewards will flip if the energy price is negative.

3.2.4 Reward 2

Notice that the reward in equation (14) will only be positive for taking a charging action if the energy price is negative. Therefore, the energy storage system will usually prioritize discharging. An alternative reward function is to formulate it in terms of the average energy price x^{avg}. We can calculate a moving average of the energy price at time step k that is a function of the past average as well as the current energy price.

$$x^{\text{avg}}[k] = (1-\beta)x^{\text{avg}}[k-1] + \beta x[k] \tag{15}$$

By including this average price in the reward function we can reward the true goal of arbitrage: to charge at low prices and discharge at high prices.

$$r_2[k] = \left(\eta_d \tilde{p}_d[k](x[k] - x^{\text{avg}}[k]) + \frac{1}{\eta_c}\tilde{p}_c[k](x^{\text{avg}}[k] - x[k])\right)\Delta t \tag{16}$$

Using this reward function, the energy storage will receive a positive reward when charging if the current energy price is lower than the average price ($x^{\text{avg}}[k] - x[k] > 0$). Similarly, the energy storage system will be positively rewarded if the current energy price is higher than the average price ($x[k] - x^{\text{avg}}[k] > 0$) and it is discharging. Note that this reward is equivalent to Reward 1 if the average price x^{avg} is equal to zero.

3.2.5 Q-Learning Algorithm

The Q-learning algorithm works by creating and updating a Q table which keeps track of the maximum expected future rewards for each action at each state. Each row in the Q table represents a state of the system and each column represents an action. The values in the Q table are an estimate of the expected reward given the state of the system indicated by the row and the action taken indicated by the column.

In the beginning, the Q table is initialized to random guesses and is not an accurate prediction of the reward based on the action and state. We use the Q-learning algorithm to update the Q table as we take actions and observe the resulting reward. Over time, the

Q table become a better and better estimate for the expected rewards and can be used to select the action with the maximum reward. The equation to update the Q table is called the Bellman equation as seen in eq (17).

$$Q[s[k], a[k]] = (1-\alpha)Q[s[k], a[k]] + \alpha(r[k] + \gamma \max_a Q[s[k+1], a]) \tag{17}$$

To avoid getting stuck with a particular action, a ϵ-greedy algorithm is used. Based on a parameter ϵ, either a random action will be taken or the action will be selected by the largest Q value in the row associated with the current state. The full Q-learning algorithm can be seen in Algorithm 2. For the simulations below, we used $\alpha = 0.4$, $\gamma = 0.2$, $\epsilon = 0.8$, and $\beta = 0.2$.

Algorithm 2 Q-Learning Algorithm

Require: Initialize e_0, γ, α, ϵ, and β if using reward 2

for $k \in 0, \cdots, K$ **do**
  Get current energy state $e[k]$ and energy price $x[k]$
  Generate random number $y \in [0, 1]$
  if $y < (1-\epsilon)$ **then**
    take random action $a[k]$
  else
    $a[k] = \max_a Q[s[k], a]$
  Calculate reward $r[k]$ using (14) or (16)
  Update $Q[s, a]$ in Q Lookup Table using eq (17)
  Calculate $e[k+1]$ using (5)

3.3 Random Action Algorithm

In order to have a baseline to compare the reinforcement algorithms above we will also implement a random algorithm. With this algorithm, for each time step the energy storage system will take a random action out of the three actions available.

Algorithm 3 Random Action Algorithm

for $k \in 0, \cdots, K$ **do**
  Select random action $a[k] \in \mathcal{A}$

4 Simulation

For implementation, we used PyCharm to interact with python code. To support the tool, we used python packages numpy and pandas to handle the data analysis and matplotlib to do the visualizations. Python is quite adept at solving reinforcement learning problems

and many online references exist [22, 23, 24]. In addition, we utilized a more built out and dedicated reinforcement python learning package, gym, to create the energy storage system environment [25, 26, 27]. In order to implement the optimization problem, we used the cvxpy package for python which is a modeling framework for solving Disciplined Convex Programming problems. We then use the Gurobi optimization solver to get a solution.

4.1 Real Time Prices

The main dataset that we used was the historical Real-Time Five-Minute LMPs from ISO New England. ISO New England (ISO NE) is an independent, non-profit grid operator for all the New England states. ISO NE ensures the day-to-day reliable operation of New England's bulk power generation and transmission system as well as oversees the administration of the region's wholesale electricity markets. One of these markets is the Real-Time Energy Market which lets market participants buy and sell wholesale electricity during the course of the operating day. Locational marginal prices (LMPs) are prices that determine the cost of supplying one additional unit of energy at a particular spot on the grid [28, 29].

ISO NE publishes the final version of the LMPs online which can be downloaded in 6 day increments [30]. Supplemental to the LMP dataset is metadata provided by ISO NE on the 1186 different nodes in the system. For this simulation, we have selected to use prices associated with Location ID 363 which is the region of Burlington, VT. The dataset read in as a pandas DataFrame is just over 400 kilobytes in the python memory with each price stored as a 64 bit float. This data set includes 25,908 observations of five minute LMPs.

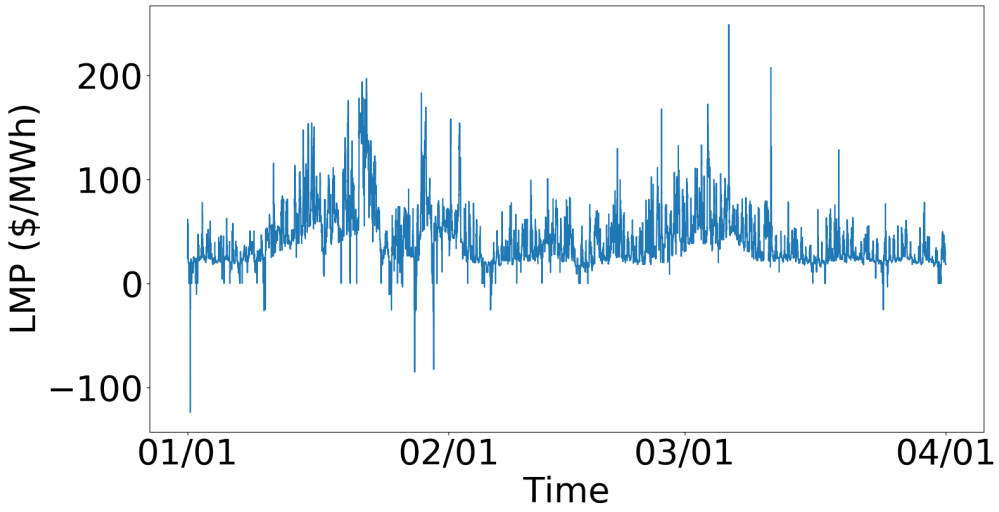

Figure 1: Timeseries Plot of ISO NE Five-Minute LMP 363 for Jan-Mar 2019

A timeseries plot of the five minute data for January through March can be seen in Figure 1. This view gives us a sense of the range of prices but since there are so many observations

it does not give detailed intuition into the distribution. Next, we look at a histogram of the data in Figure 2. Here we can gain some knowledge that the majority of prices are between 20 and 50 $\frac{\$}{MWh}$. A similar intuition can be gained from the probability density function estimation plot seen in Figure 3.

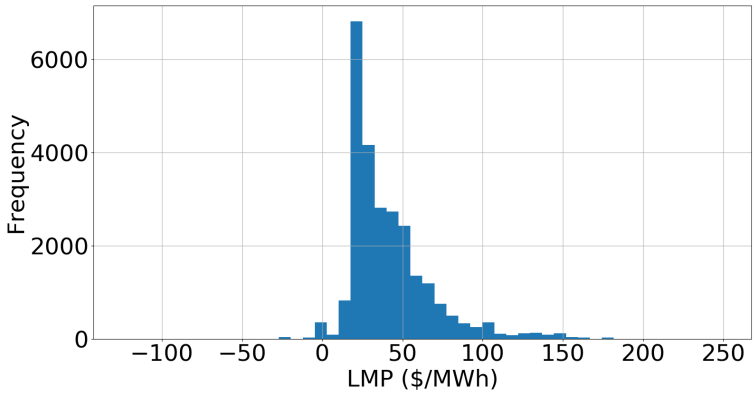

Figure 2: Histogram Plot of ISO NE Five-Minute LMP 363 for Jan-Mar 2019

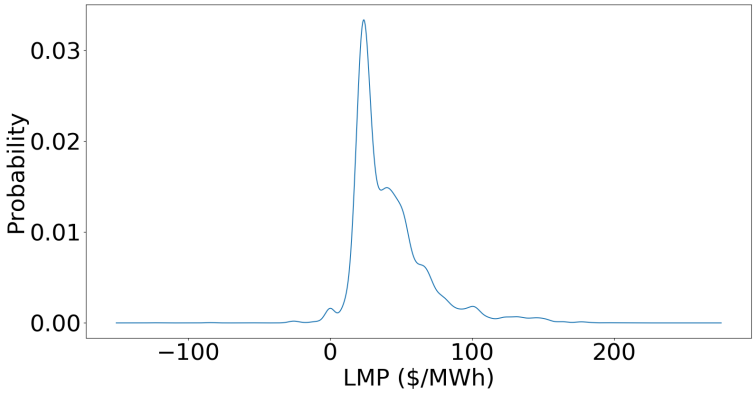

Figure 3: Probability Density Function Plot of ISO NE Five-Minute LMP 363 for Jan-Mar 2019

However, the graphics in Figure 2 and Figure 3 do not give good knowledge about the magnitude and occurrence of any outliers in the data. In order to see this better, we recreate the histogram plot with a log Y axis scale as seen in Figure 4. With the current data set the lowest prices are around -125 and highest prices are 200-250 but these only occur a handful of times in the two months. Our intuitions of the data can be confirmed by looking at some statistical measures in Table 1.

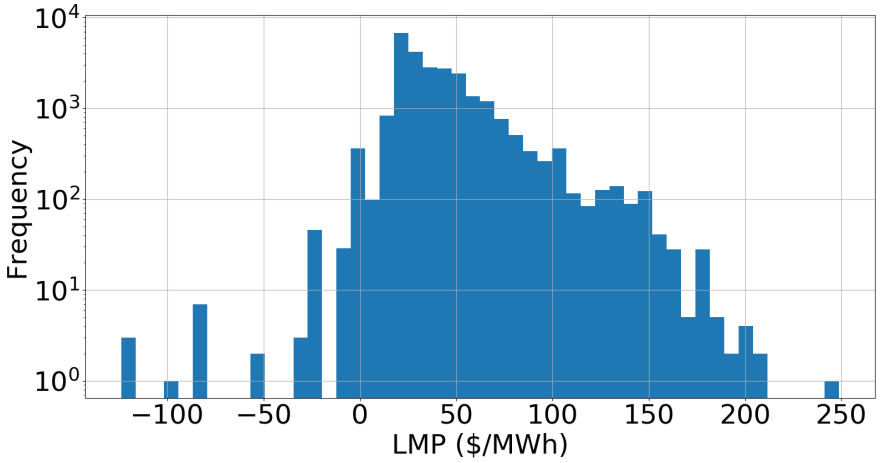

Figure 4: Log Histogram Plot of ISO NE Five-Minute LMP 363 Dataset for Jan-Mar 2019

count	mean	std	min	25 %	50%	75%	max
25908.0	41.6	26.63	-124.04	23.69	34.63	51.85	248.84

Table 1: Summary Statistics of ISO NE Five-Minute LMP Dataset for Jan-Mar 2019

4.2 Energy Storage System

For the simulations, a 20 MWh energy storage system was used with charging and discharging maximum power of 5 MW. This means that we have installed a energy storage system with 4 hour discharging or charging capability. We assumed that the charging and discharging efficiencies are 0.9 and self-discharge rate is 0.1% of the current energy per hour.

4.3 Simulations Results

The total cost for each algorithm is shown in Table 2. For the non-deterministic methods (random and Q learning), the simulation was ran 20 times and the total costs were averaged.

	Optimal Solution	Random Action	Q (Reward 1)	Q (Reward 2)
Total Cost ($)	-57,799	37,370	-20,303	-13,571

Table 2: Average energy cost for Jan-Mar using four control algorithms

As you can see, the optimization control scheme is able to make a large profit by performing arbitrage (charging when the electricity price is cheap and discharging when it is high), however, it has perfect forecast of the electricity prices which is not realistic for a real time control strategy. By contrast, the random action strategy ends with a new cost due to buying energy when it is expensive as well as loosing energy due to the inefficiencies of

charging and discharging. Simulating the Q learning strategies proved that they could operate the energy storage system at a new profit without any future knowledge of the energy prices. Based on the 20 simulations, operating an energy storage system using reward 1 is on average more profitable that using reward 2.

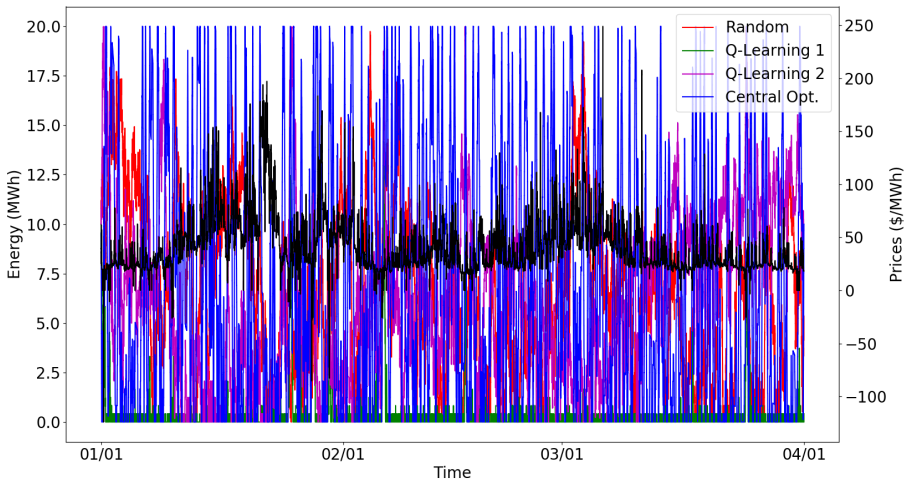

Figure 5: Energy of energy storage system for Feb and Mar using four control algorithms

We can take a closer look at what is happening by looking at the energy and power for each 5 minute time step over the two month period looked at. In Figure 5, the energy of the energy storage system is plotted for each time step for each of the three control algorithms for a single simulation. The random algorithm has random behavior and tends to stay around a 50% charge. The optimization algorithm has much more drastic changes as it will charge at maximum power while energy is cheap and discharge at maximum power while energy is expensive to maximize profits. The Q learning algorithm using reward 1 has a unique and noteworthy behavior. Since the reward is only positive when the energy storage system is discharging, the Q learning arbitrage policy will always want to chose to discharge when it can. Therefore, what we end up seeing is it discharges quickly and stays near empty for the whole three months, only charging when prices are low enough. This was the original impetus behind using reward 2. By designing the reward around the average price it reward charging as well as discharging. Using reward 2, the energy storage has a much higher energy on average.

We can also take a look at the distribution of charging and discharging prices for each control algorithm. These can be seen in Figure 6 and the log histogram can be seen in Figure 7 where positive power indicates charging and negative indicates discharging. As previously stated the Random algorithm has a wider distribution of powers and is more likely to charge or discharge at full power than the other algorithms. The optimization algorithm always knows the perfect times to charge and discharge and ends up charging or discharging less than the maximum power in order to maximize the energy arbitrage profit. The Q learning control with reward 1 will discharge as much as possible and often stay at

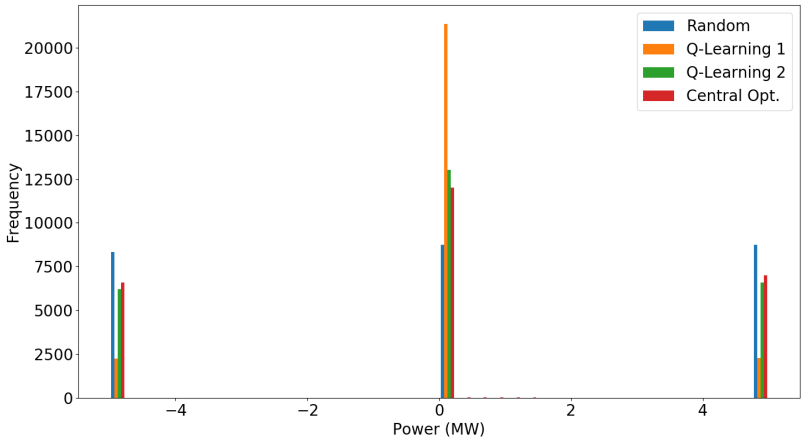

Figure 6: Power Histogram of energy storage system for Feb and Mar using four control algorithms

completely empty hence the higher frequency of power equal to zero. Figure 7 shows a similar story but highlights the charging and discharging power that is less than the maximum rate.

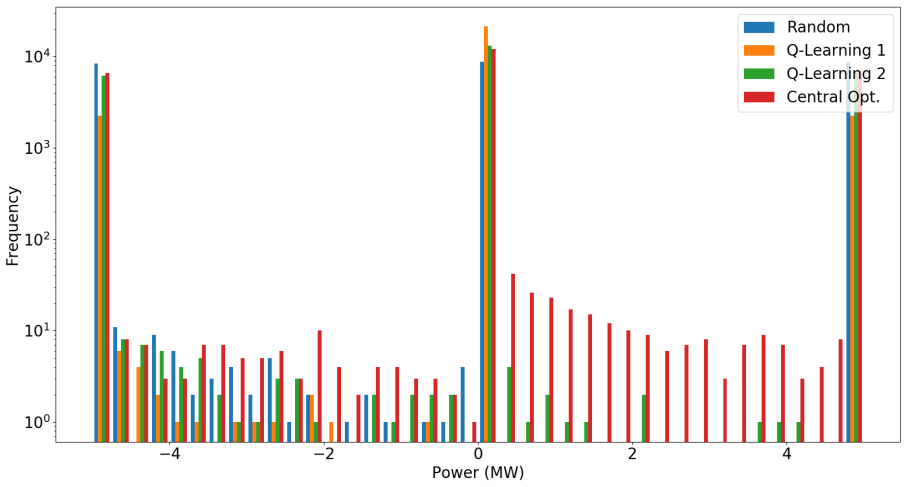

Figure 7: Power Log Histogram of energy storage system for Feb and Mar using four control algorithms

A snapshot of the energy storage system control can be seen in Figure 8. The bottom graph shows the real time energy price while the top two show the power and energy at each time step. The optimization control can be seen to not charge or discharge and then charge (positive power) when the location marginal prices are low and discharging (negative power) when they are high. The Q learning algorithms also tend to do this but with more random motion and not always at the optimal time.

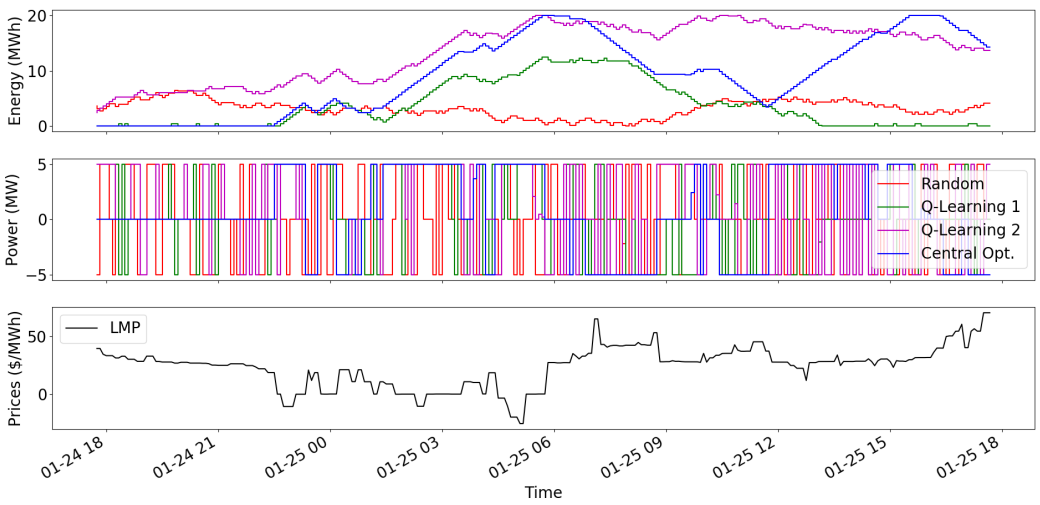

Figure 8: 24 Hour Sample of energy storage system control using four control algorithms

Finally, we can we view how the cost of controlling the energy storage system accumulates over time. Figure 9 shows the cumulative profit for each time step from January to March. This was calculated by using a cumulative sum on the negative of the cost function from eq (9) and is a function of the energy storage system's charging and dischagring power. The perfect forecast optimization algorithm has a steady increase in profits whereas the Q learning strategy tends to have more sudden jumps in profits. Contrasting the optimization approach, the random algorithm has a steady decrease in profits or a steady increase in costs over time.

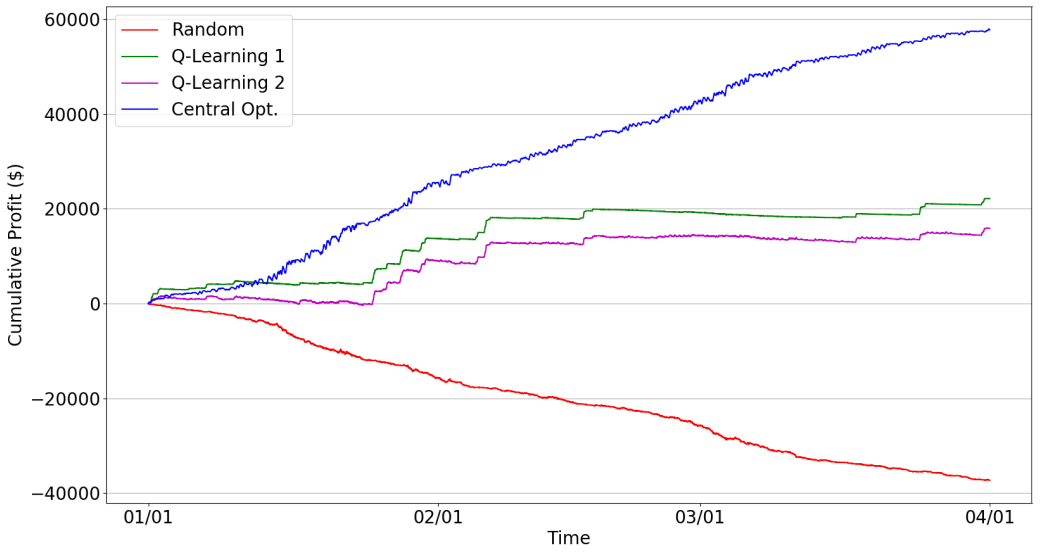

Figure 9: Cumulative Profit of four control algorithms

5 Conclusion

In this paper, we derived an effective arbitrage strategy for energy storage system control in fast real-time markets via reinforcement learning. First, we carefully define the state and action spaces, as well as the reward function in the Q-learning algorithm such that the objective of the reinforcement learning algorithm coincides with our goal of maximizing the profit through energy arbitrage. We were successful in showing that a Q-learning policy can control the charge and discharge behavior of the energy storage system to achieve a profit. Simulating this algorithm demonstrated that the strategy performs much superior than taking a random action while not achieving the optimal profit as calculated by a perfect forecast optimization problem.

Future work could involve further tuning of the Q learning parameters and sensitive analysis of the parameters. Also, future research could compare more adaptations of reinforcement learning algorithms such as incorporating multi-time steps rewards and a neural network which is referred to as deep Q learning. In addition, the reward function could be adapted to incorporate other considerations such as battery degradation or other value streams for energy storage such as frequency regulation or backup power.

References

[1] J. Eyer, G. Corey, and S. N. Laboratories, *Energy Storage for the Electricity Grid: Benefits and Market Potential Assessment Guide : a Study for the DOE Energy Storage Systems Program*. SAND (Series) (Albuquerque, N.M.), Sandia National Laboratories, 2010.

[2] R. H. Byrne and C. A. Silva-Monroy, "Estimating the maximum potential revenue for grid connected electricity storage: Arbitrage and regulation," 2012.

[3] C. J. C. H. Watkins, "Learning from delayed rewards," 1989.

[4] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279–292, May 1992.

[5] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, pp. 9–44, Aug 1988.

[6] G. A. Rummery and M. Niranjan, *On-line Q-learning using connectionist systems*, vol. 37. University of Cambridge, Department of Engineering Cambridge, England, 1994.

[7] R. S. Sutton and A. G. Barto, *Reinforcement learning - an introduction*. Adaptive computation and machine learning, MIT Press, 1998.

[8] H. Wang and B. Zhang, "Energy storage arbitrage in real-time markets via reinforcement learning," *CoRR*, vol. abs/1711.03127, 2017.

[9] C. Guan, X. Lin, Y. Wang, , S. Nazarian, and M. Pedram, "Reinforcement learning-based control of residential energy storage systems for electric bill minimization," in *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 637–642, Jan 2015.

[10] E. Kuznetsova, Y.-F. Li, C. Ruiz, E. Zio, G. Ault, and K. Bell, "Reinforcement learning for microgrid energy management," *Energy*, vol. 59, pp. 133 – 146, 2013.

[11] G. Henri and N. Lu, "A multi-agent shared machine learning approach for real-time battery operation mode prediction and control," in *2018 IEEE Power Energy Society General Meeting (PESGM)*, pp. 1–5, Aug 2018.

[12] A. S. Zamzam, B. Yang, and N. D. Sidiropoulos, "Energy Storage Management via Deep Q-Networks," *arXiv e-prints*, p. arXiv:1903.11107, Mar 2019.

[13] L. Xiao, X. Xiao, C. Dai, M. Peng, L. Wang, and H. V. Poor, "Reinforcement learning-based energy trading for microgrids," *CoRR*, vol. abs/1801.06285, 2018.

[14] T. Matiisen, "Demystifying deep reinforcement learning." https://neuro.cs.ut.ee/demystifying-deep-reinforcement-learning/.

[15] S. Kim and H. Lim, "Reinforcement learning based energy management algorithm for smart energy buildings," *Energies*, vol. 11, p. 2010, 08 2018.

[16] B. Kim, Y. Zhang, M. van der Schaar, and J. Lee, "Dynamic pricing and energy consumption scheduling with reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 7, pp. 2187–2198, Sep. 2016.

[17] K. Rahbar, J. Xu, and R. Zhang, "Real-time energy storage management for renewable integration in microgrid: An off-line optimization approach," *IEEE Transactions on Smart Grid*, vol. 6, pp. 124–134, Jan 2015.

[18] T. Hubert and S. Grijalva, "Modeling for residential electricity optimization in dynamic pricing environments," *IEEE Transactions on Smart Grid*, vol. 3, pp. 2224–2231, Dec 2012.

[19] J. Qin, Y. Chow, J. Yang, and R. Rajagopal, "Online modified greedy algorithm for storage control under uncertainty," *IEEE Transactions on Power Systems*, vol. 31, pp. 1729–1743, May 2016.

[20] Y. Xu and L. Tong, "Optimal operation and economic value of energy storage at consumer locations," *IEEE Transactions on Automatic Control*, vol. 62, pp. 792–807, Feb 2017.

[21] D. Mears, H. Gotschall, H. Kamath, E. P. R. Institute, U. S. D. of Energy, T. I. (Firm), and E. P. Corporation, *EPRI-DOE Handbook of Energy Storage for Transmission and Distribution Applications*. EPRI, 2003.

[22] ADL, "An introduction to q-learning: reinforcement learning," Sep 2018. https://medium.freecodecamp.org/an-introduction-to-q-learning-reinforcement-learning-14ac0b4493cc.

[23] A. Juliani, "Simple reinforcement learning with tensorflow part 0: Q-learning with tables and neural networks," Aug 2016. https://medium.com/emergent-future/simple-reinforcement-learning-with-tensorflow-part-0-q-learning-with-tables-and-neural-networks-d195264329d0.

[24] V. Kurama, "Reinforcement learning with python," Nov 2018. https://towardsdatascience.com/reinforcement-learning-with-python-8ef0242a2fa2.

[25] OpenAI, "A toolkit for developing and comparing reinforcement learning algorithms." https://gym.openai.com/docs/.

[26] G. Hayes, "Getting started with reinforcement learning and open ai gym," Feb 2019. https://towardsdatascience.com/getting-started-with-reinforcement-learning-and-open-ai-gym-c289aca874f.

[27] A. Poddar, "Making a custom environment in gym," Jul 2018. https://medium.com/@apoddar573/making-your-own-custom-environment-in-gym-c3b65ff8cdaa.

[28] T. Orfanogianni and G. Gross, "A general formulation for lmp evaluation," *IEEE Transactions on Power Systems*, vol. 22, pp. 1163–1173, Aug 2007.

[29] I. N. England, "Faqs: Locational marginal pricing." https://www.iso-ne.com/participate/support/faq/lmp.

[30] I. N. England, "Pricing reports." https://www.iso-ne.com/isoexpress/web/reports/pricing/-/tree/lmps-rt-five-minute-final.