AA 228 Final Project:
Maximizing Profit from Battery Operation on the Grid
Given Future Price Uncertainty

Antonio Aguilar, Justin Appleby, Robert Spragg

December 7th, 2018

Abstract

In this project, we apply model-free reinforcement learning algorithms to the problem of energy arbitrage; we aim to maximize profit for independent battery operator who buys, holds and sells energy from the CAISO grid at favorable times. We determine a theoretical upper bound on performance in the case of perfect knowledge of future prices, replicate a Q-Learning benchmark and test more complex learning strategies.

1 Introduction and Motivation

Energy storage is critical to a future electricity system in which renewable energy sources account for an increasingly large share of the power mix. High renewable penetration in places like California has caused energy supply and pricing to become more volatile, thus increasing the need for load-shifting entities and the opportunity for them to profit while performing that work.

Our objective is therefore to develop strategies to learn profitable real-time policies for a battery engaging in energy arbitrage at two nodes in the California Independent System Operator Grid (CAISO). The main challenge is price uncertainty; energy prices are settled in real time every 5 minutes and are notoriously hard to model accurately given their wide variability and the resolution of available data.

Following a recent paper by Wang and Zhang at the University of Washington [1], we choose to employ model-free reinforcement learning in the form of Q-Learning. This allows us to avoid explicitly assuming a price distribution while keeping us flexible to operate under changing and non-stationary prices.

In our work, we expand upon their model to include more complex exploration strategies such as ϵ-greedy and softmax. We evaluate the expansion the state space of Q-Learning in order to account for features such as hour and day of week. We also analyze the performance benefits of different price discretizations. Finally, we validate our resulting best policy with data from a second year. Our results show that proper discretization of prices has a large impact on profit, and that the epsilon-greedy method algorithm performs best when paired with an expanded state space.

2 Methods and Modeling

We first settled on a battery model. Typically, grid-scale energy storage projects are designed to cycle (completely discharge at the highest rate) in 3 or 4 hours [2]. We select a cycle time of 3 hours. The specs of our simulated battery are the following:

- Capacity: 30 MWh
- Max Charge Rate: 10 MW
- Max Discharge Rate: 10 MW

Then, to evaluate the performance of our learning algorithms, we established both a naive baseline and a theoretical upper bound for performance.

2.1 Optimal solution

In a deterministic setting (i.e. future prices are known), energy arbitrage is a straightforward linear optimization problem that has been well-studied [3]. The problem can be framed as follows:

$$\begin{aligned} \text{max: } & \sum_{t=1}^{\tau} p_t \left(\eta_d d_t - \frac{1}{\eta_c} c_t \right) \\ & E_t = E_{t-1} + c_t - d_t \\ \text{subject to: } & E_{\min} \leq E_t \leq E_{\max} \\ & 0 \leq c_t \leq C_{\max} \\ & 0 \leq d_t \leq D_{\max} \\ \text{with variables: } & c_t, d_t, t \end{aligned} \quad , \forall t \in \tau$$

2.2 Naive baseline

We then developed a naive policy, given our understanding of the net load and average prices on the CAISO energy market, as shown in Figures 5 and 6. The policy is described below:

- Charge at maximum charging rate from 2am to 5am
- Discharge at maximum discharge rate from 7am to 10am
- Charge at maximum charging rate from 10am to 1pm
- Discharge at maximum discharge rate from 7pm to 10pm

The monthly earnings of the optimal and naive policies are compared in Table 1 and in Figure 7. From this figure, we can see that the naive policy makes steady gains throughout the month of August 2017, but suffers major profit hits at several points. These are due to charging when price is very high.

2.3 Energy Arbitrage as a Markov Decision Process

2.3.1 State Space

We define the state space chiefly along the dimensions of price and energy level. We discretize the price into $m = 100$ bins. Three binning strategies are used: quantile cuts, even cuts, and

Profit Comparison, August 2017

Policy	Los Altos
Deterministic Linear Optimization	**$77,742.36**
Naive Policy 1 (Nighttime Charging)	**$9,959.10**

Table 1: Score (profit) comparison for various policies

"smart cuts", in which prices below $100/MWh are cut into 80 quantile bins, while prices above $100/MWh are split into 20 quantile bins. We discretize the battery charge level into $n = 36$, given that our battery can charge or discharge 1/36 of its maximum capacity in the 5-min observation intervals.

We also expand upon the state-space to include several features suspected to have an impact on price dynamics. The following were tried separately and evaluated on a full year of real-time prices from two nodes – one in Los Altos, CA and one in Fresno, CA.

- Day of week: $\in \{0, ..., 6\}$
- Weekday/weekend: $\in \{0, 1\}$
- Peak/off-peak (3pm - 8pm on weekdays is considered "peak"): $\in \{0, 1\}$
- Hour: $\in \{0, ..., 23\}$

2.3.2 Action Space

From the formulation of the linear optimization, the following 3 things hold true, regardless of the price signal:

Lemma 1 *The optimal charge and discharge pro les $\{c_t^*, d_t^*, \forall t \in \mathcal{T}\}$ satisfy:*

1. *At least one of c_t^* or d_t^* is 0 at any time t;*
2. $c_t^* = \{0, \min\{C_{\max}, E_{\max} - E_{t-1}\}\}$
3. $d_t^* = \{0, \min\{D_{\max}, E_{t-1} - E_{\min}\}\}$

The action space therefore contains no more than three actions: charge at full throttle, hold idle, or discharge at full throttle.

$$\mathcal{A} = \{-\tilde{D}_{\max}, 0, \tilde{C}_{\max}\}$$

2.3.3 Reward Function

Intuition suggests that the reward function used to dictate the relative value of different states should be the money earned or spent in taking that action, so that the utility accurately reflects the profits made. However, Wang and Zhang [1] found that this method leads to under-exploring of the state-space because this reward function penalizing for charging at any price. They made significant gains by employing a reward function that considers a moving average of recently seen prices:

$$r_t = \begin{cases} (\overline{p}_t - p_t)\tilde{C}_{\max} & \text{if charge} \\ 0, & \text{if hold} \\ (p_t - \overline{p}_t)\tilde{D}_{\max}, & \text{if discharge} \end{cases}$$

where average price $\overline{p}_t$ is given by the following equation, with η being a smoothing parameter we set at 0.9:

$$\overline{p}_t = (1 - \eta)\overline{p}_{t-1} + \eta p_t,$$

With this reward formulation, the search algorithm will try to explore by charging when the current price is below the moving average.

3 Results

First, we evaluate our exploration and discretization strategies. Figure 1 shows the cumulative profits made using a number of combinations. (The global optimum is shown in the same figure for reference; Softmax was evaluated but not included due to consistently poor performance.) Without knowing future prices, Q-learning can obtain about 50% of the theoretical maximum for the year.

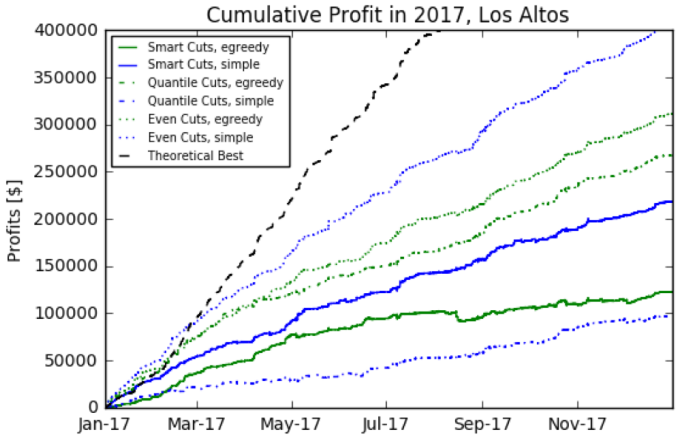

Figure 1: Cumulative profit using various online learning strategies — Los Altos, 2017

While the simple exploration strategy prevails over ϵ-greedy above, we select ϵ-greedy when expanding the state space due to the increased need for exploration. We chose to conduct the same procedure on two different nodes, to see if having different price signals has an effect on performance. Figure 2 shows the cumulative profits of various state-space expansions in the year 2017 at Los Altos and Fresno using ϵ-greedy. State spaces with larger domains, hour and day-of-week, prevail over those with binary domains.

3.1 Validation

To examine the robustness of the epsilon-greedy, expanded-state-space algorithm to different scenarios, validation was performed by training on one year and running the derived policy offline during a different year. The policy was also re-run offline for its own year to determine amount of over-fitting that occurs. The data is shown in Figure 3.

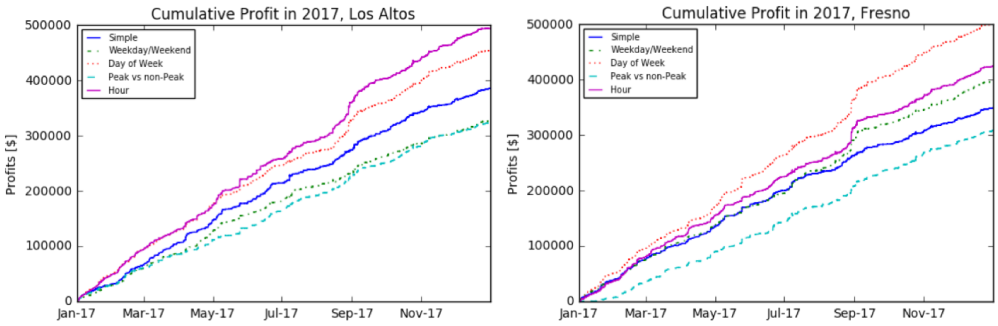

Figure 2: Cumulative profit from online learning using various state spaces and epsilon-greedy — Los Altos (left), Fresno (right). 2017

		Trained On	
		2016	2017
Evaluated On	2016	$76,607	$54,305
	2017	$69,997	$103,289

Figure 3: Validation of epsilon-greedy algorithm. As expected, the offline policy performed better on the year it was trained on.

3.1.1 Model Sensitivity

Models such as epsilon-greedy have inherent variability in their results, since the action known to be best at a given time is taken with probability ϵ, and the remaining other actions are taken with even distribution summing to $1-\epsilon$. Therefore, we calculated the distribution of profit for one month for the epsilon-greedy and softmax algorithms, simulating each 200 times. The results are shown in Figure 4 below. The findings in Figure 1 are validated by the histogram, which reveals that ϵ-greedy performs better. Of concern is the high variance observed. Future work might be well served to analyze techniques to reduce it.

4 Conclusion

Various learning algorithms were applied to power price data at two nodes in the CAISO grid. For each algorithm, the state space was varied, to find the features that contribute most to learning how to perform battery arbitrage. The results show that the ϵ-greedy algorithm typically performed best when the state space includes additional features. The feature that provided the largest increase in profit compared to the naive baseline was hour of day. Finally, policies trained during a given year always performed less well on a different year, but still performed better than the naive policy.

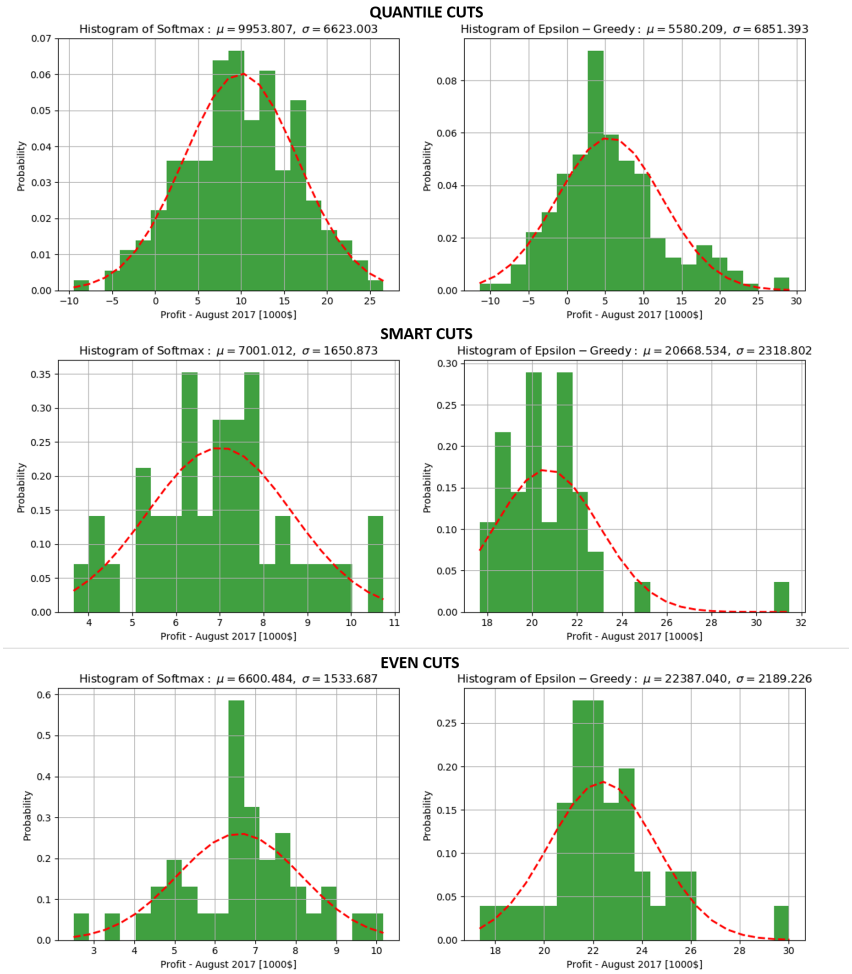

Figure 4: Distribution of profit for softmax (left) and epsilon-greedy (right) using various discretization techniques — Los Altos, August 2017.

5 Contributions

Antonio

- Set up original q-learning algorithm
- Set up battery simulator

Justin

- Scraped data from CAISO OASIS API
- Implemented epsilon-greedy algorithm

Robert

- Solved global optimum benchmark (using CVX)
- Added model states for day-of-week, hour, weekday vs weekend

References

[1] H. Wang and B. Zhang, "Energy storage arbitrage in real-time markets via reinforcement learning," *IEEE Power and Energy Society General Meeting*, February 2018.

[2] "Svce signs major contracts for california's largest solar-plus-storage projects," October 2018.

[3] A. et. al., "Cost-optimization of battery sizing and operation," *eCAL - UC Berkeley*, May 2016.

6 Appendix

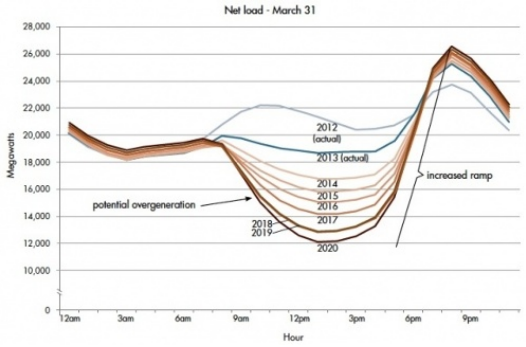

Figure 5: CAISO net load (March 31) — Note the change in net load that has occurred in the past decade due to increased solar penetration.

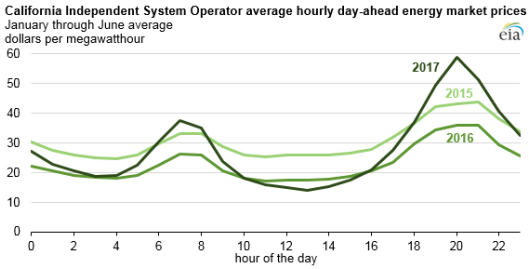

Figure 6: CAISO average hourly day-ahead prices — Note the changes that have occured due to increased solar penetration.

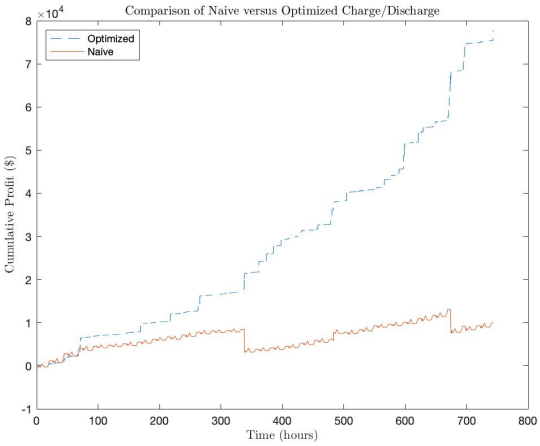

Figure 7: Comparison of the cumulative reward for August 2017 using the optimal (deterministic) solution and the naive policy at the Los Altos node.

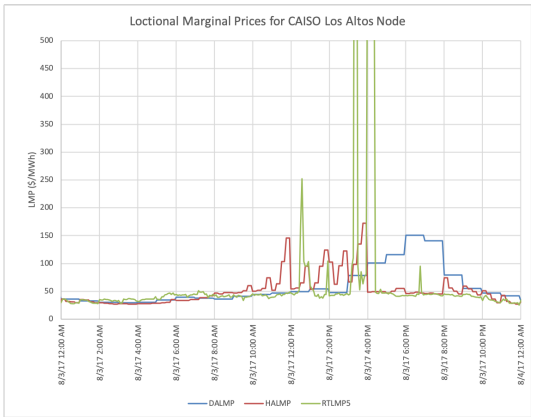

Figure 8: Sample of real-time, hour-ahead and day-ahead electricity LMPs at CAISO's Los Altos node. The day-ahead market is settled at 10am the day before.