Energy Storage Price Arbitrage in Real Market Simulations via Deep Q-Learning Networks

Eliza Cohn, Neal Ma
Earth and Environmental Engineering
Columbia University
New York, USA
{ec3766, nam2252}@columbia.edu

Davit Shadunts
Electrical Engineering
Columbia University
New York, USA
ds4355@columbia.edu

***Abstract*—The increased integration of renewable energy resources into the power grid necessitates energy storage units to mitigate the instability of their power generation. Energy storage arbitrage is the charging and discharging strategy that maximizes this revenue. However, determining the optimal mechanism to govern their behavior is an open research question. Reinforcement learning is quickly becoming a popular tool to solve this problem due to its flexibility and ability to encode historical data. This paper develops a deep Q-learning (DQN) agent that represents an energy storage unit participating in the electricity market. Through real-world market simulations, it is demonstrated that the DQN agent with the highest performance is the one that always bids slightly below the real-time price of electricity.**

***Index Terms*—Reinforcement learning, deep Q-learning, energy storage, power system economics, bidding, market power**

I. Introduction

The electricity market in the United States has seen a rapid transformation over the last two decades. Originally established as vertically integrated entity over a century ago, recent years have witnessed the deregulation of the electric system to mitigate inefficiencies caused from government monopolies. This development was finalized at the turn of the century, when electrical energy was first officially considered a commodity and markets were created for trading. The benefits of this unbounding are the promotion of competition, increased system transparency, and a fair market in the sense that no system operator will favor one generator over another [1].

The structure of the electricity market facilitates a vast array of different types of generators that can enter in their bids for energy generation. In addition, the physical constraints of the power grid requires that supply has to meet demand at all time. In order to achieve this, system operators decide which generators to dispatch to meet demand at each time step [2]. Each generation unit submits a bid that represents the amount of deliverable power and their marginal cost of generation. The bids are aggregated by the system operator and cleared in merit order from lowest price to highest. The clearing price is defined by the point where supply and demand meet. Once this is determined, all generators that are cleared to produce energy will receive the clearing price [3].

Decarbonizing energy systems through a cleaner power grid is essential for combating climate change. Decarbonization has demanded an increase in energy storage systems to mitigate the instability associated with variable renewable energy resources (VRE) [4]. Energy storage systems (ESS) play a critical role in providing energy to the power grid when VRE output is low. Additionally, energy storage systems can generate revenue via arbitrage by charging at low prices and discharging at high prices. The optimal mechanism for ESS bidding strategies to maximize both revenue and system reliability is currently an open research question.

Mixed integer linear programming (MILP) was one of the first techniques proposed to solve energy storage arbitrage. Krishnamurthy et al. provide an updated framework for this, focusing on incorporating the stochasticity of energy price into their optimization model [5]. However, MILP algorithms face the canonical curse of dimensionality when increasing problem complexity and also an inability to handle non-linear relationships without introducing approximations. Stochastic dynamic programming (SDP) is another approach for determining optimal arbitrage strategies. It has a flexible architecture for incorporating uncertainty in electricity price demonstrates improved results when benchmarked against MILP [6].

Reinforcement learning is quickly becoming one of the most popular methods for energy storage arbitrage. A major advancement of RL over typical optimization techniques is the enablement of offline learning; decoupling the expensive learning portion from the real-time decision process. Q-Learning algorithms are a simple but effective solution choice by discretizing the possible bidding action space. Formulations can vary from the point of view of the system operator or the generator, but designers typically try to encode historical price information in the reward function and use Monte Carlo simulations for testing case studies [7], [8]. Deep Q-Learning (DQN) is another option, and is typically applied to represent generator agents. DQN is preferred over Q-Learning due to its ability to handle larger and more complex state spaces that more accurately represent the dynamics of the electricity market [9]. Multi-agent learning is emerging as popular strategy by modeling more of players in the electricity market [10].

However, no matter the complexity of the RL technique used, almost the entire body of literature is dedicated to modeling uncertainty in electricity price, VRE output, and consumer demand. This paper seeks to address that gap in the literature by expanding the uncertainty considerations into the electricity market itself. Our model presents a novel and lightweight methodology for representing the uncertainty of a bid being accepted into the market during both training and evaluation. DQN was the selected RL algorithm due to its ability to learn complex state spaces while maintaining an efficient implementation. The contributions of this paper are:

1) A DQN algorithm is used to learn the optimal bidding behavior of an energy storage system.
2) A novel method for representing the uncertainty of bids being accepted in training allows for the DQN agent to target conservative, neutral, or risky strategies.
3) A realistic market simulation is created for evaluation that directly computes the bidding curve of the participating generators. [1]

The paper is organized as follows. Section II presents the methods implemented such as the design of the market simulation, the bidding uncertainty, and the DQN network. Section III presents the results and Section IV provides a discussion of insights observed and concludes the paper.

II. Methods

A. Energy Storage Formulation

Energy storage arbitrage is a relatively well-studied problem. A battery can be simply defined as:

- E : Storage capacity (MWh), the total amount of energy that can be stored in the battery
- P_{max} : Power capacity (MW), the maximum charge and discharge rate of the battery as seen by the grid operator
- η : Charge/discharge efficiency of the battery
- SOC : State-of-Charge (unitless), in $[0, 1]$ indicating the fraction of the total storage capacity, E, that a battery currently is charged to

Of these four variables, only the SOC evolved through time while the rest act as constraints and time evolution factors:

$$\begin{aligned} SOC_t =& SOC_{t-1} + (p_t \eta - b_t/\eta)/E \\ & 0 \leq b_t \leq P_{max} \\ & 0 \leq p_t \leq P_{max} \\ & p_t \times b_t = 0 \\ & SOC_t \in [0, 1] \; \forall \; t \end{aligned}$$

Here, p_t is the charging power at time t and b_t is the discharging power (positive) at time t. Because simultaneous charging and discharging is disallowed, at least one of p_t and b_t must be zero at each time step.

This simplified battery model does not account for other important factors such as fixed discharge costs, resting draw, and non-linear charge/discharge efficiencies at different SOC levels.

B. Market Simulation

The market simulation is built to emulate the real-time market as closely as possible while minimizing computational complexity. Due to the high complexity of electricity markets and many simplifications made to rapidly develop this simulator, not all inaccuracies and deviations from real market operations are detailed here.

[1][Code Available]: https://github.com/nmadev/EnergyStorageRLBidder

1) Theory of Market Clearing: At each timestep in the real-time market, all bidders participate via submitting price bids to charge/discharge at full power capacity at that timestep. Typically, energy storage units will bid for both charging and discharging, i.e. the battery operator will submit a minimum price at which they will fully discharge at as well as a maximum price at which they will fully charge at.

Once these bids are collected, the system operator determines whether the grid needs more power (positive demand/batteries will be dispatched the discharge) or the grid has excess power (negative demand/batteries will be dispatched to charge). Market clearing, as displayed in Fig. 1, follows the same procedure for both charging and discharging with slight modifications.

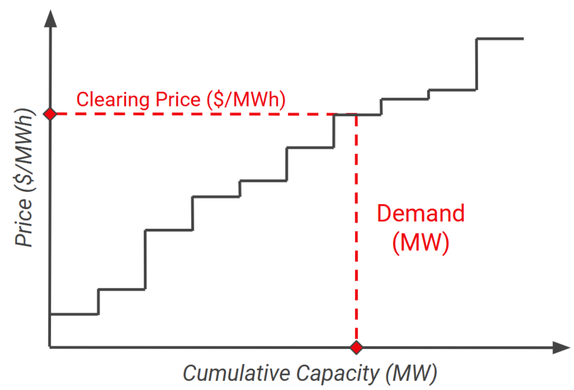

Fig. 1. Market clearing example for positive demand

Bids are sorted in merit order, or increasing in price, if demand is positive and in reverse merit order, or decreasing in price, if demand is negative. The demand at the timestep is then used to determine the clearing price. All of the bids below the demand at the timestep, including the bid that fully satisfies demand, are cleared in order. The clearing price is set at the price of the bid that fully clears the demand. All cleared bids receive the clearing price for charging/discharging at that time. All bids above the clearing bid are not cleared, the corresponding batteries do not charge/discharge.

If the battery that submitted the clearing bid only needs to partially charge/discharge to satisfy demand, that battery will only be cleared partially. If multiple bids are submitted at the clearing price, some bids will be selected at random to be cleared. One important distinction to note is the difference between real-time price and clearing price. In the real market, real-time price is set as the clearing price at all times. In this simulation, the real-time price is taken as the bid of honest bidders, as described in section II-C. The market is then simulated with these bids and bids from trained bidders to determine a clearing price. This choice was intentional to allow market interaction by the

Once this clearing process has finished, the appropriate power and corresponding cost/revenue is returned to each energy storage unit and the market simulation is stepped to the next timestep.

2) Data Integration: Historical energy data is obtained [11] by scraping California Independent System Operator (CAISO) Daily Energy Storage Reports [12]. This includes hourly and sub-hourly data for real-time price, day-ahead price, energy storage awards, and bid distributions among many other energy storage-related fields. From this comprehensive dataset, the real-time price and energy storage awards data are used. The former is used both to train the bidders and to directly determine bids for "honest" bidders, as detailed in the next section. The energy storage awards, or the historical cleared and dispatched battery capacity, is used in the market simulator to determine the amount of capacity cleared in the market. These data are also averaged on a time-of-day basis to be used in the clearing probability function for training each bidder, explained in greater detail in sections II-D2 and II-D3.

C. Honest Bidder

Honest bidders are created to simulate market participation by conventional thermal generators. In electricity markets, thermal generators are typically required to bid at their marginal cost and cannot arbitrarily raise and lower their bids like energy storage units can. The capacity of these generators typically heavily outweigh the capacity bid by energy storage and are much larger factors in determining the clearing price at any given time. However, as the interconnected capacity of energy storage units increase sharply, the proportion of the market made up by honest bidders decreases.

Honest bidders are simulated as energy storage units that always submit bids, λ_t^b, exactly at the real-time price, λ_t^{RTP}:

$$\lambda_t^b = \lambda_t^{RTP}$$

The higher the proportion of total market capacity that is taken up by these bidders, the closer the real-time price and market clearing price are expected to be.

D. Deep Q-Learning Bidder

The Deep Q-Learning Bidder optimizes the revenue of energy storage arbitrage systems by simulating various bidding strategies within a real-time electricity market environment. The state space includes Real-Time Price (RTP), State of Charge (SOC), and the timestamp. The action space consists of a set of discrete bidding multipliers $a_t \in \{0, 0.1, ..., 2.0\}$, which scale the RTP to determine the bid price:

$$\text{bid}_t = a_t \times \text{RTP}_t$$

The environment behaves as an offline market simulator where the probability of bid acceptance is influenced by the current bid. The reward function is designed as follows:

$$r_t = \begin{cases} \text{power} \times \text{clearing price} \times \text{granularity} & \text{Bid accepted} \\ 0 & \text{Bid not accepted} \end{cases}$$

1) DQN Network: The DQN network consists of the following components:

- **Input:** State features: RTP, SOC, and the timestamp.
- **Hidden Layers:** Two fully connected layers with 64 neurons each, using ReLU activation function.
- **Output:** Q-values for all possible bidding actions.

The `cleared_action` function introduces bidding uncertainty, enabling the DQN agent to adapt its behavior based on conservative, neutral, or risky strategies. The full implementation, including state updates, reward calculations, and Q-network optimization is presented in Algorithm 1.

Algorithm 1 DQN for Energy Storage Bidding in RT Market

Input: Learning rate lr, discount factor γ, batch size $batchsize$, episodes $episodes$, replay buffer $buffer$, market data $data$, ϵ-decay schedule
Output: Optimized Q-network ω
Initialize Q-network ω and replay buffer $buffer$
Initialize action space $A = \{0.0, 0.1, \ldots, 2.0\}$
for each episode $e = 1, 2, \ldots, episodes$ **do**
  $s_0 \leftarrow (\text{RTP}, 0.5, \text{ts})$ ▷ SOC initialized to 0.5
  Initialize cumulative reward $rsum \leftarrow 0$
  for each time step $t = 1, 2, \ldots, T$ **do**
    Compute ϵ using decay schedule
    Sample random value $r \sim \mathcal{U}(0, 1)$
    if $r > \epsilon$ **then**
      Select action $a_t = \arg\max_a Q(s_t, a; \omega)$
    else
      Select random action $a_t \sim A$
    end if
    Compute bid $bid_t = \text{RTP} \cdot a_t$
    Compute probability of bid acceptance:
      $is\_cleared =$
        $cleared\_action(\text{RTP}, bid_t, attitude, \text{SOC}, ts)$
    if $is\_cleared = 1$ **then**
      $power \leftarrow \min\left(\frac{\text{SOC}\cdot\text{capacity}}{\text{granularity}}, \text{power\_max}\right)$
      $\text{RTP} \leftarrow \frac{\text{RTP}}{\text{eff}}$
    else if $is\_cleared = -1$ **then**
      $power \leftarrow -\min\left(\frac{(1-\text{SOC})\cdot\text{capacity}}{\text{granularity}}, \text{power\_max}\right)$
      $\text{RTP} \leftarrow \text{RTP} \cdot \text{eff}$
    else
      $power \leftarrow 0$
    end if
    Execute action a_t
    Observe reward $r_t = power \cdot \text{RTP} \cdot \text{granularity}$
    Compute next state $s_{t+1} = (\text{RTP}, \text{SOC}_{t+1}, \text{ts}_{t+1})$
    Store transition (s_t, a_t, r_t, s_{t+1}) in $buffer$
    if size of $buffer > batchsize$ **then**
      Sample minibatch of $batchsize$ from $buffer$
      for each (s_i, a_i, r_i, s_i') in minibatch **do**
        Compute target:
        $y_i = r_i + \gamma \max_a Q(s_i', a; \omega)$
        Update Q-network ω to minimize loss:
        $$L(\omega) = \frac{1}{batchsize} \sum_{i=1}^{batchsize} \big(y_i - Q(s_i, a_i; \omega)\big)^2$$
      end for
    end if
    Update cumulative reward $rsum \leftarrow rsum + r_t$
    Update state $s_t \leftarrow s_{t+1}$
  end for
  Print $rsum$ for episode e
end for

2) Static Bidding: To simulate the uncertainty of a bid being accepted in the training environment, the market dynamics were approximated using the `cleared_action` function. An overview of this mechanism is given in Figure 2.

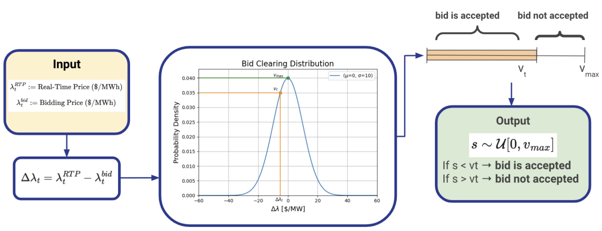

Fig. 2. Overview of the Bidding Uncertainty Mechanism implemented in the `cleared_action` function

This lightweight implementation is more scalable for the training environment compared to generating the bidding curve for each time step in evaluation. For each time step, the real-time price (λ_t^{RTP}) and the proposed bid (λ_t^b) are inputs. The deviation from the RTP is computed as $\Delta\lambda_t$. The main engine of this mechanism is a normal distribution $\mathcal{N}(\mu, \sigma)$. We then compute $v_t = \mathcal{N}(\Delta\lambda_t, \mu, \sigma)$, the value used as a threshold for acceptance in a uniform distribution $\mathcal{U}[0, v_{max}]$. The maximum value of the normal distribution is given by:

$$v_{max} = \frac{1}{\sqrt{2\pi\sigma^2}} \tag{1}$$

By generating a random sample $s \sim \mathcal{U}[0, v_{max}]$ and checking if it is above or below v_t, we can simulate a bid being accepted or not with uncertainty. The closer the bid is to the mean of the distribution, the more likely it is to be accepted. Different *attitudes* can be adopted by centering the normal distribution above, below, or around the real-time price. These attitudes are defined respectively as *risky, conservative* and *neutral*. The term *static bidding* refers to the the distribution means remaining constant for each attitude. The distributions for each attitude are given in Figure 3.

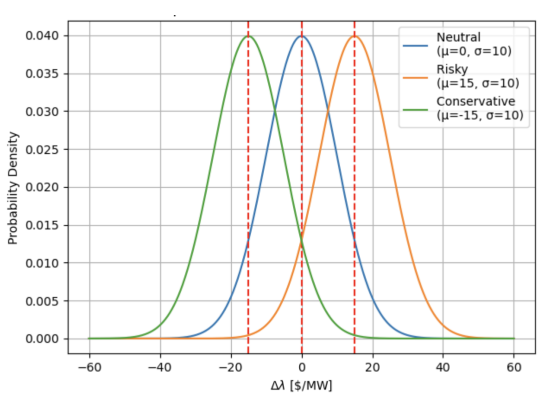

Fig. 3. Normal Distributions of Static Attitudes

Once it is determined if the bid is accepted or not, the historical average demand profile is used to inform the bidder if they have been cleared for charge or discharge. Given the fact the regional demand is predictable and consistent over a diurnal period, we say that the bidder is cleared for discharge when $d_t > 0$ and cleared to charge when $d_t < 0$. The demand profile is given in Figure 4.

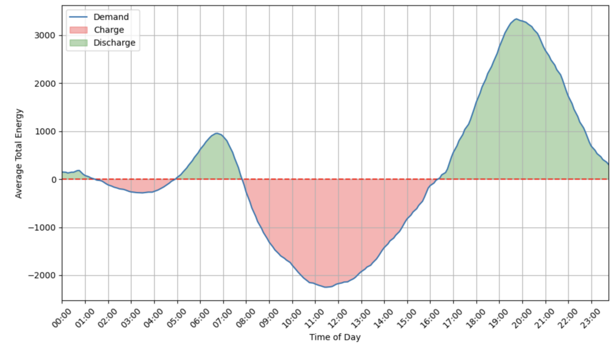

Fig. 4. Average Smoothed Daily Demand Profile

3) Dynamic Bidding: In order to take advantage of the demand profile being known, the *dynamic bidders* were developed as an extension of the static bidders. The dynamic bidders shift the mean of their acceptance distribution based on the demand, as shown in Figure 5.

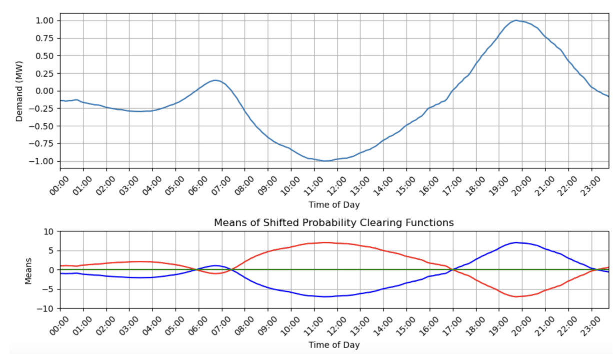

Fig. 5. Mean Shifting under Dynamic Attitudes

The conservative bidder bids above the RTP when demand is low and bids below the RTP when demand is high. This follows the logic that the bids are more likely to be accepted in periods of high demand when the bid is slightly below the RTP and vice versa for periods of low demand. The risky bidder enacts the opposite of this behavior, and the neutral bidder remains centered at the RTP for all time steps.

The `cleared_action` function encodes these steps by returning 1 if the bid has been accepted to discharge, -1 if the bid has been accepted to charge, and 0 if the bid was not accepted.

E. Experimental Design

For this study, two parameters are varied between tests: i) the proportion of market capacity constituted by honest bidders and ii) the scaled capacity of the maximum demand with respect to the total market capacity.

The first parameter is varied to test how each bidding strategy will perform as more energy storage capacity interconnects to the market. These proportions are $\{0.0, 0.25, 0.5, 0.75, 0.9\}$. The lower the proportion of honest bidders, the higher the proportion of battery capacity in the grid.

The second parameter is varied to include/exclude certain bidders from the market and test the performance under those conditions. These scale factors are $\{0.8, 1.0, 1.2\}$. A

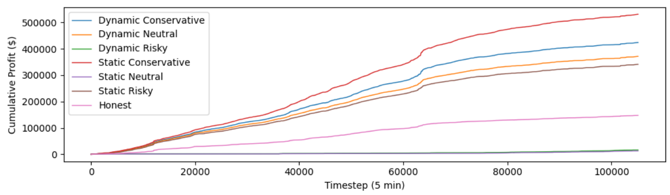

Fig. 6. Cumulative profit of each bidding algorithm over 1 year of simulations under 75% honest bidders and no demand scaling

higher proportion indicates more bids will be cleared for both charging and discharging and is expected to allow for greater profits for risky behavior. On the other hand, a lower proportion would likely very frequently exclude these risky bidders from the market.

III. Results

Due to the volume of data output from each simulation under each parameter, only select results are detailed below.

Fig. 6 shows the cumulative profit obtained by each type of bidder in the year-long simulation. In this simulation, demand is not scaled at all and 75% of the total market capacity is made up by honest bidders.

Here, the best performing bidding algorithms are those trained with both dynamic and static *conservative* probability clearing functions with near $500000 in profits each. The next two best performing algorithms are dynamic neutral and static risky, respectively, each reaching just under $400000 in profit. The average honest bidder profit is much lower at approximately $200000 and finally, dynamic risky and static neutral bidders are the worst performing at only approximately $30000 each. The following ranking is the same order of earned profits across all simulations:

1) Static Conservative
2) Dynamic Conservative
3) Dynamic Neutral
4) Static Risky
5) Honest (If in the simulation)
6) Dynamic Risky
7) Static Neutral

Interestingly, this ordering holds exactly for energy cycled, as displayed in Fig. 7.

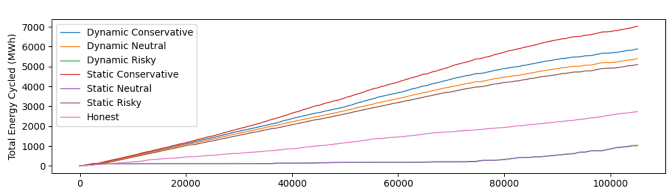

Fig. 7. Total energy cycled through each bidder over 1 year of simulations under 75% honest bidders and no demand scaling

This result indicates that the more a battery participates in a market, the more profit that battery will make over time. This result of greater dispatch leading to greater profit is further analyzed in Section IV from a game-theory perspective.

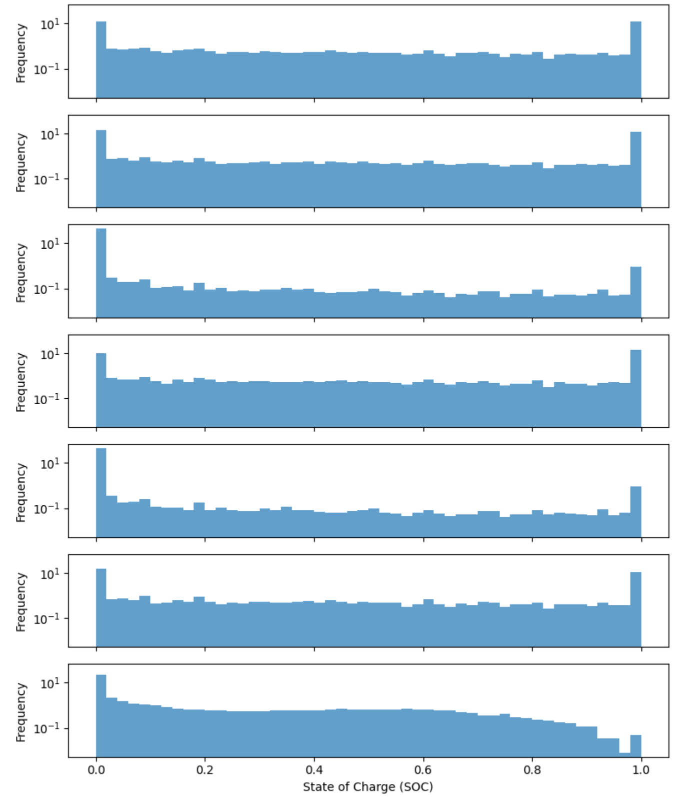

Fig. 8. State-of-Charge Distributions for all bidders with 90% honest bidder composition and demand scaled to 120%. Bidding strategies top to bottom: dynamic conservative, dynamic neutral, dynamic risky, static conservative, static neutral, static risky, honest.

The energy cycling of these batteries can also be studied through the distribution of SOC values throughout the simulation, as displayed in Fig. 8.

Another important factor of integrating energy storage into the grid is total system cost. This is calculated as the total profit made by all bidders in the system over the entire year and normalized to the total capacity of the system. Fig. 9 shows how this varies over all demand scaling values and honest bidder proportions.

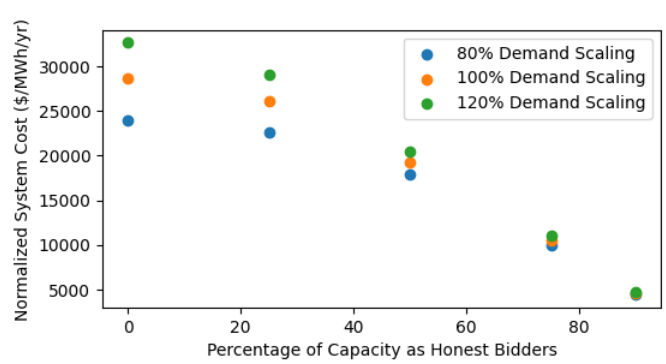

Fig. 9. Normalized system cost in each simulation scenario.

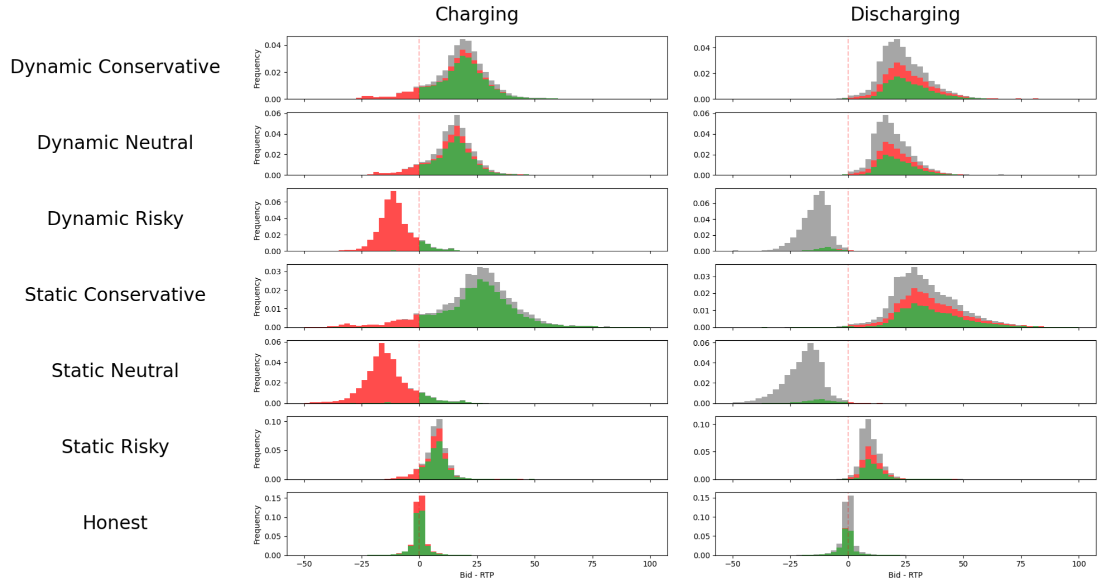

Fig. 10. Bid distributions for charging (left) and discharging (right) times with accepted bids (green), rejected bids (red), and bids when battery capacity is not available to charge/discharge (gray).

Over every scenario, as the proportion of honest bidders increase, the normalized system cost decreases. This is due to the honest bidder, which on-average is one of the least profitable bidders, constituting larger portions of the overall profit, an unsurprising result. Similarly, as demand is scaled up, more energy storage bidders are allowed to participate in the market at a greater frequency often driving up the price and actually increasing the overall normalized system costs. This is likely due to larger participation allowing for higher bids to set the clearing discharge prices. These results are both unsurprising but are still good confirmation that the market is operating in an expected manner.

Perhaps most interestingly, bid price distributions for each individual bidder can be visualized as in Fig. 10. These distributions further confirm that greater market participation, as seen in the dynamic and static conservative bidders, the greater the profit. A very odd result are the distributions of dynamic risky and static neutral charge/discharge distributions. One would expect both the charge and discharge distributions of static risky to be centered around $\lambda_t^b = \lambda_t^{RTP}$ instead of being negative as seen here. Additionally for the dynamic risky behavior, the charging bids being below the no difference value is expected, but the discharge bids following a similar behavior does not make sense.

IV. Discussion & Conclusion

A game theoretic framework can provide insights as to why the static conservative bidder performs the best across all simulations. This framework is similar to the mechanism of a *first price auction* (FPA), in which the highest bidder wins and pays the price they bid [13]. In this case, all bidders are paid the clearing price, which is highest bidding price accepted.

The Nash equilibrium shows that the optimal bidding strategy β in a first price auction with n participants is known to be:

$$\beta(v_i) = \frac{n-1}{n} v_i \tag{2}$$

Where v_i is the independent valuation of the bidder. Here, the real-time price is the valuation of energy shared by all bidders. From the equilibrium in a FPA, we know that the optimal strategy is to bid slightly below this valuation. This maps to the strategy of the conservative bidder, which is to always bid slightly below the real-time price no matter the demand or time of day. In the case of the market simulation presented, the conservative bidders follow this strategy and see that their bids are the most likely to be accepted. Simply put, the more the bids are accepted, and therefore are able to maximize revenue.

The paper presents a deep Q-learning based agent that represents an energy storage system participating in the real-time electricity market. Through market simulations, it is show that the the agent trained with the static conservative bidding strategy performs the best.

Future work would determine if the conservative bidder still performs the under enhanced battery modeling, training environment, and market simulations. For battery modeling, a more accurate physical representation would model ambient discharge and fixed cost integration. In the training environment, using a discounted reward function may encourage more strategic charging behavior over a longer horizon. The day-ahead-price of electricity could also be tested as an additional state variable and input into the DQN, since this forecast is available for real-world market participants. Determining how these agents impact market behavior would be very informative for real-world operations. If the agents are price influencers, then their impact on the price of electricity should be included in the model. Finally, testing the transferability of the trained agents to other electricity markets would determine the robustness of learned bidding behavior.

REFERENCES

[1] D. S. Kirschen and G. Strbac, *Fundamentals of power system economics*, 2nd ed. Nashville, TN: John Wiley & Sons, Sep. 2018.

[2] RFF, "Us electricity markets 101," 2022. [Online]. Available: https://www.rff.org/publications/explainers/us-electricity-markets-101/

[3] ISONE, "How resources are selected and prices are set in the wholesale energy markets," 2024. [Online]. Available: https://www.iso-ne.com/about/what-we-do/how-resources-are-selected-and-prices-are-set

[4] N. Zheng, X. Qin, D. Wu, G. Murtaugh, and B. Xu, "Energy storage state-of-charge market model," *Electrical Engineering and Systems Science*, Jul. 2022.

[5] D. Krishnamurthy, C. Uckun, Z. Zhou, P. R. Thimmapuram, and A. Botterud, "Energy storage arbitrage under day-ahead and real-time price uncertainty," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 84–93, Jan. 2018.

[6] N. Zheng, J. Jaworski, and B. Xu, "Arbitraging variable efficiency energy storage using analytical stochastic dynamic programming," *IEEE Trans. Power Syst.*, vol. 37, no. 6, pp. 4785–4795, Nov. 2022.

[7] F. Meng, Y. Bai, and J. Jin, "An advanced real-time dispatching strategy for a distributed energy system based on the reinforcement learning algorithm," *Renew. Energy*, vol. 178, pp. 13–24, Nov. 2021.

[8] H. Wang and B. Zhang, "Energy storage arbitrage in real-time markets via reinforcement learning," in *2018 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, Aug. 2018.

[9] T. Zonjee and S. S. Torbaghan, "Energy storage arbitrage in day-ahead electricity market using deep reinforcement learning," in *2023 IEEE Belgrade PowerTech*. IEEE, Jun. 2023, pp. 1–7.

[10] B.-C. Lai, W.-Y. Chiu, and Y.-P. Tsai, "Multiagent reinforcement learning for community energy management to mitigate peak rebounds under renewable energy uncertainty," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 3, pp. 568–579, Jun. 2022.

[11] nmadev, "Caiso-energystorage," 2024. [Online]. Available: https://github.com/nmadev/CAISO-EnergyStorage

[12] CAISO, "Caiso daily energy storage reports," 2024. [Online]. Available: https://www.caiso.com/library/daily-energy-storage-reports

[13] Greenwald and Oyakawa, "First-price sealed-bid auctions," 2017. [Online]. Available: https://cs.brown.edu/courses/cs1951k/lectures/2020/first_price_auctions.pdf