Optimizing Energy Carbon Emissions and Cost for Grid-Interactive Efficient Buildings with Q-Learning

Anna Edmonds
Stanford University
edmondsa@stanford.edu

Bradley Hu
Stanford University
bradleyh@stanford.edu

Jenna Mansueto
Stanford University
mansueto@stanford.edu

Abstract

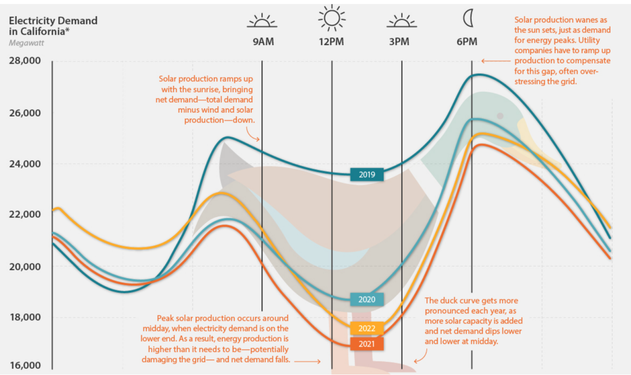
The buildings in the United States account for 38% of the nation's energy consumption. Therefore, it's vital for these buildings to enhance their energy efficiency to minimize greenhouse gas emissions through dynamically using their battery storage systems and strategically importing and exporting energy at specific times of the day. To effectively manage these energy systems in buildings, reinforcement learning can be applied as it adapts to real-time external signals and is capable of modeling sequential decisions as the carbon intensity, demand, supply, price, changes throughout the day. In this paper, we explore the application of Q-learning to this problem and test the effects of varying rewards function on building and district-level carbon emissions, electricity cost and energy consumption. Although Q-learning remains constrained by its requirement of discretized state and action spaces, we find that custom reward functions can still improve agent performance.

1. Introduction

1.1. Background and Motivation

In the United States, the residential and commercial sectors accounted for about 21% and 17% respectively—38% combined—of total U.S. energy consumption in 2022 [1] and contributing 30% to greenhouse gas emissions [2]. In pursuit of zero-emission buildings, strategies are being developed that include the electrification of building utilities and transitioning the power grid toward renewable energy sources like solar and wind [3]. However, this shift toward electrification in buildings may lead to increased demands on the current electrical grid and without concurrent efforts to decarbonize power generation, this transition could counterintuitively result in higher greenhouse gas emissions.

The primary challenge in integrating renewable energy sources, such as solar energy, is the mismatch between supply and demand, highlighted by phenomena like California's duck curve [4]. The curve illustrates the challenges in managing solar energy influx when the grid experiences reduced demand when there is higher generation of renewable energy, but increased demand during lower generation of renewable energy. The issue is complicated by the variability of renewable energy supply, dependent on weather conditions, and demand fluctuations based on local usage patterns.



Solar production ramps up with the sunrise, bringing net demand—total demand minus wind and solar production—down.

Solar production wanes as the sun sets, just as demand for energy peaks. Utility companies have to ramp up production to compensate for this gap, often over-stressing the grid.

Peak solar production occurs around midday, when electricity demand is on the lower end. As a result, energy production is higher than it needs to be—potentially damaging the grid— and net demand falls.

The duck curve gets more pronounced each year, as more solar capacity is added and net demand dips lower and lower at midday.

Figure 1. Duck Curve in California Shows Peak Demand is Mismatched with Peak Production of Solar Panels

To mitigate the effects of the duck curve, strategies like load shedding and load shifting are critical. Load shedding involves reducing electricity use during peak demand times, whereas load shifting adjusts the timing of electricity use to align with periods of lower demand, promoting the use of more affordable and cleaner energy. The usage of load shifting and load shedding is further optimized when each building uses their own battery storage and has their own solar panel, but still can use electricity from the grid. If a building has its own battery and self-generation then it can interact with the grid and export it's renewable energy and not have to solely rely on the grid. In literature, these types of buildings are referred to as grid interactive efficient buildings (GEBs) where they can provide energy during times of supply deficit, where it attempts to coordinate selling and purchasing of energy in order to minimize energy bills and greenhouse gas emissions [5].The correct application of GEBs could further help provide grid resilience in the face of power outages, this is an ongoing research project by the US General Services Administration [6].

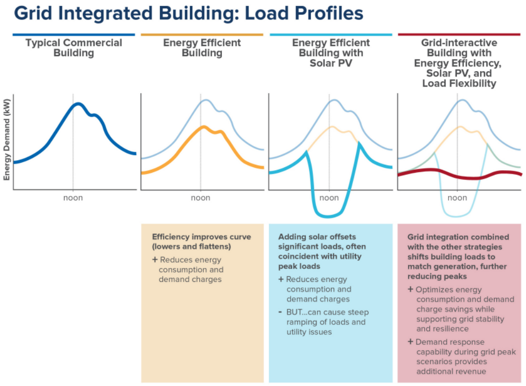


Figure 2. Grid Interactive Efficient Buildings with Energy Efficiency, Solar Panels and Storage

1.2. Problem Definition

The battery storage and building energy demand needs to be carefully managed simultaneously in many buildings to unlock their full energy efficiency potential and reduce the cost. However, at present, the effectiveness of demand response by GEBs is limited due to the requirement for manual input at real time. The carbon intensity and cost of energy fluctuate based on real-time factors like weather conditions and neighborhood demand. To maximize the reduction of greenhouse gas emissions and cost of energy by effectively utilizing battery storage in GEBs and grid energy at optimal times, a sophisticated control architecture is crucial that functions at real time.

1.3. Approach to Managing GEBs

Reinforcement Learning (RL) emerges as an ideal solution for dynamic energy management, in the context of Grid-Interactive Efficient Buildings. RL's adaptability to real-time external signals allows it to efficiently manage the use of renewable energy through a process of sequential decision-making. Throughout the day, RL systems gather real-time data on variables such as the carbon intensity of energy, overall energy consumption, and solar power generation. This continuous data influx enables the system to refine its strategies for battery charging and discharging, as well as for determining optimal periods to either draw from or supply energy to the grid.

This complex decision-making process is effectively modeled using a Markov Decision Process (MDP). In this framework, the energy system of a building acts as the decision-making agent. It evaluates various actions, such as determining the percentage to charge or discharge battery storage, at each time interval (typically hourly) based on the current energy system state. This state encompasses various factors like energy prices, carbon intensity, demand, and supply.

The foundation of this model is the Markov assumption, which states that the future state of the system depends solely on its current state and the action chosen by the agent. Leveraging this assumption, the RL system progressively learns and improves its decision-making, aiming to optimize energy usage and reduce costs. This ongoing learning process is key to enhancing the building's energy efficiency and sustainability, adapting to changing conditions while minimizing environmental impact.

1.4. Related Work

Researchers are exploring a variety of algorithms to address the complex challenge of dynamic energy allocation in Grid-Interactive Efficient Buildings.

To facilitate this, studies are utilizing CityLearn, an OpenAI Gym environment to simulate the interactions of buildings interacting with the dynamic grid. CityLearn enables the creation and benchmarking of advanced algorithms specifically tailored for demand response research. CityLearn is widely recognized for its application in various studies, including incentive-based and collaborative demand response, coordinated energy management, benchmarking of demand response algorithms, and voltage regulation. CityLearn operates independently of external grid signals, leveraging active storage systems for load shifting. It includes detailed models of buildings, electric heaters, heat pumps, thermal storage, batteries, and photovoltaic systems (PVs). CityLearn simulates 17 buildings, each equipped with a battery and a PV system, aligning with real-world community building specifications. Control agents in CityLearn manage the State of Charge (SOC) of batteries, determining energy storage or release at each timestep. The platform ensures building loads are met based on predetermined data and includes a backup controller to maintain system constraints like occupant comfort and base load requirements. This controller also regulates battery charging and discharging to meet these load requirements [7].

There are a few papers who have used CityLearn to simulate this dynamic environment in the pursuit of making buildings more energy efficient. A recent research paper published this year propose a MERLIN framework which is multiagent offline and transfer learning framework that uses both CityLearn environment to model demand response by controlling batteries and smart meter data from 17 net zero single family homes. This paper demonstrates that MERLIN, combined with a CityLearn digital twin, effectively maintains comfort, caters to unique occupant behavior, and achieves building and district-level objectives, showing comparable performance with limited data training and effective policy transfer across different buildings. [8]. Another research paper attempts to use a multi-agent PPO learning approach to schedule the powering of the buildings using CityLearn and the dataset from the 2022 NeurIPS Challenge. This paper builds three modules, including feature engineering module, forecasting module and reinforcement learning module [9]. We will be taking a different RL approach, but with the same data set than the research paper.

2. Methodology

2.1. Data

For this study, we utilized the CityLearn Challenge 2022 dataset (https://www.aicrowd.com/challenges/neurips-2022-citylearn-challenge), which contains time series data for 17 buildings in Fontana, California from August 1, 2016 to July 31, 2017. These buildings are a part of a Zero Net Energy community, and each is equiped with a 6.5 kWh capacity home battery as well as 4 kW or 5 kW photovoltaic capacity. In addition to the data collected from these buildings, the dataset has been augmented with weather data from the Los Angeles International Airport Weather Station [10], electricity price data from the community's utlities provider [11], and carbon intensity data from the Electric Power Research Institute.

2.2. Environment Design

We used this dataset in a CityLearn environment, which allowed us to easily experiment with different agent and environment configurations.

To constrain the scope of the project, we limited our analysis to a subset of the buildings and observations. We randomly selected two buildings as well as a one week time period, using a random seed for reproducibility. Although the original dataset identifies observations by hour, day type, and month, we chose to limit our observation space to hour alone given our time and computational constraints. Similarly, we limit our action space to charging/discharging electrical storage (battery) only, and discretize this continuous action space ($[-1.0, 1.0]$) into 12 bins.

Hyperparameter	Value
ϵ	1
ϵ_{decay}	0.0096
ϵ_{min}	0.01
γ	0.98
i	10
α	0.0001

Table 1. Hyperparameters.

2.3. Data Preprocessing

In addition to this filtering and discretizing of the data, we also used min-max normalization for the carbon intensity values, as the original range of values made using this field in our reward function challenging.

2.4. Baseline

We compare our agent's performance to a baseline model where buildings have no batteries, and thus no ability to intelligently store or discharge electricity.

2.5. Reinforcement Learning Agent Design

We chose to treat this problem as a discrete MDP, using a model-free approach, Q-Learning, to find an optimal policy. Q-Learning incrementally approximates the value function $Q(s, a)$ which maps state-action pairs to their associated value. To do this, it employs the following update rule, with discount factor γ and learning rate α:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s,a) \right] \tag{1}$$

We adopt an epsilon-greedy exploration strategy with epsilon-decay. We then train our agent for $\frac{m \times n \times i}{t}$ episodes, where i is an integer and t is the number of time steps per episode. A complete list of our hyperparameters can be found in Table 1.

Although CityLearn provides the option to run centralized or decentralized agent simulations (i.e. a single agent controls all buildings or each building is controlled by an independent agent), we chose to focus on the centralized case.

2.6. Reward Function Design

We implemented a primitive reward function (R_P) and two custom reward functions to explore their effects on carbon emissions, electricity cost and energy consumption. The primitive reward function is defined as the sum of negative electricity consumption (E) across all buildings, and is the default reward function for Q-Learning in CityLearn:

$$r_P = \sum_{i=0}^{n} (p_i \times -E_i) \tag{2}$$

The two custom reward functions contain a penalty term, p, which penalizes the agent for importing from the grid when a building's battery is full as well as for exporting when the battery is empty. The first reward function (R_C)is based primarily on the cost of electricity (C), while the second reward function (R_O) is based on the carbon intensity of grid power (O). The cost-based reward for a particular state of buildings 0 through n is defined as follows:

$$r_C = \sum_{i=0}^{n} (p_i \times |C_i|) \tag{3}$$

$$p_i = -(1 + sign(C_i) \times SOC_i^{battery}) \tag{4}$$

where the penalty term p_i is defined in terms of the battery level (SOC) of building i and the sign of the cost (C_i), to indicate whether energy is being imported or exported.

The carbon intensity-based reward function uses the same penalty term p, and is defined conditionally follows:

When $C_i >= 0$:

$$r_O = \sum_{i=0}^{n} (p_i \times O_i) \tag{5}$$

When $C_i < 0$:

$$r_O = \sum_{i=0}^{n} (p_i \times (1 - |O_i|)) \tag{6}$$

This conditional structure aims to encourage the agent to import from the grid when carbon intensity is low, and export to the grid when it is high.

2.7. Key Performance Indicators

We track three key performance indicators (KPIs): energy consumption, cost, and carbon emissions. These KPIs are calculated for each building individually, as well as averaged across all buildings to highlight district-level metrics. The building-level KPIs are defined as follows:

I. Energy consumption:

$$\text{electricity consumption}_i = \sum_{h=0}^{n-1} \max\left(0, E_h^i\right) \tag{7}$$

where E_h^i is the electricity consumption of building i at hour h. The total electricity consumption is thus defined as the sum of consumption across the simulation period.

II. Cost:

$$\text{cost} = \sum_{h=0}^{n-1} \max\left(0, E_h^i \times T_h\right) \tag{8}$$

where T_h is the electricity rate at hour h. The cost is thus defined as the sum of the cost of imported electricity for a given building.

III. Carbon emissions:

$$\text{carbon emissions} = \sum_{h=0}^{n-1} \max\left(0, E_h^i \times O_h\right) \tag{9}$$

where O_h is the carbon intensity of power from the grid at hour h. We thus define carbon emissions as the sum of carbon emissions due to energy imports for a given building.

For the remainder of this paper, we will focus on district-level averages as we are most concerned with minimizing cost and emissions at the level of neighborhoods.

3. Results and Discussion

Due to the limitations of tabular Q-learning and and the constrained state and observation space in general, the results of Q-learning do not beat those of a baseline no-battery model. However, in both cases of our custom reward functions, the results of Q-learning showed an improvement over that of the primitive CityLearn reward function implementation across all three KPIs. This can be seen in our results in Figures 3, 4, and 5.

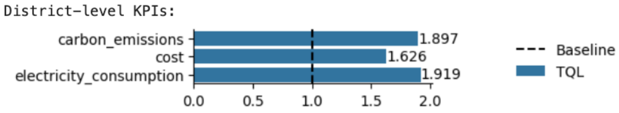


Figure 3. District KPIs for Agent Trained on Primitive Reward

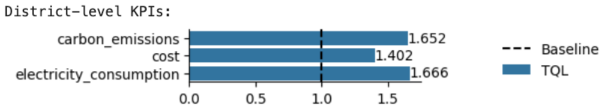


Figure 4. District KPIs for Agent Trained on Cost Minimized Reward

We theorize that this is due to the penalty term, p, in the custom reward functions, which encourages the agent to appropriately charge and discharge the battery. Interestingly, both custom reward functions performed similarly across all KPIs, suggesting that the inclusion of cost vs. emissions-related parameters had little impact on the intended outcomes (in other words, incorporating

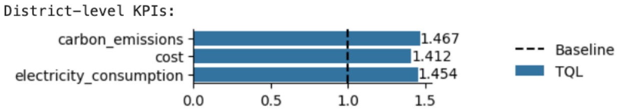


Figure 5. District KPIs for Agent Trained on Carbon Intensity Minimized Reward

carbon-intensity did not significantly impact the carbon emissions KPI).

Additionally, we evaluated the performance of our Q-learning algorithm in terms of exploration and exploitation.

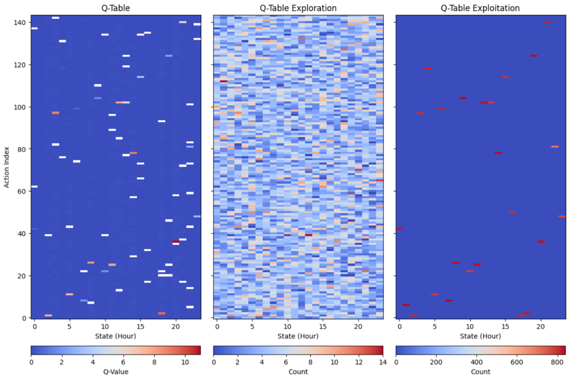


Figure 6. Q-Table Showing Exploration vs. Exploitation

From the left table of Figure 6, we see that the Q-values for most action indices are similar and take on a low value. The middle table shows that virtually all state-action pairs have been visited at least once, meaning that our random exploration was successful. The right figure shows how many state-action pairs were exploited. For each state, only one action was ever an exploitation candidate, and given the considerable exploration done by the agent, this was likely the best candidate. These tables suggest that the Q-learning agent did find the optimal policy for our abstracted problem, suggesting that perhaps the abstraction itself is responsible for poor performance.

Additionally, from visualizing the Battery SOC profiles of the two buildings (Figure 7), we see that the Q-learning algorithm seems to adopt a repetitive charge and discharge pattern across the simulation period. This is likely due to the fact that our only active observation is *hour*, and thus the agent does not learn to vary behavior for different day types.

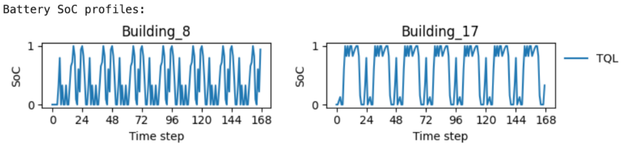


Figure 7. Battery SOC

3.1. Limitations

As suggested in the previous suggestion, our results were hindered by the constraints imposed by the Q-learning algorithm. As we had to discretize the state and action space, we inevitably lost aspects of the problem in this abstraction. The algorithm itself suffers from the curse of dimensionality, in that the Q-table grows exponentially with increases in the state and action space – thus, while adding more nuanced observations such as day-type or increasing the number of action bins could result in improved performance, we would quickly become limited by the time and compute required. Further, the adoption of Reinforcement Learning (RL) in building controls faces three key barriers: 1) the increasing time and data requirements for training as the number of buildings expands, making the process more complex; 2) the need for control systems to be both secure and robust, ensuring stability and protection against external threats; and 3) the challenge in achieving generalizability of RL controllers across diverse buildings, balancing customization with adaptability. These issues underscore the necessity for advanced and efficient RL solutions in building management.

3.2. Future Work

Faced with the limitations of Q-Learning, particularly the curse of dimensionality, our current algorithm falls short in fully optimizing the energy systems of grid-interactive efficient buildings. To address this, exploring alternative Reinforcement Learning algorithms is essential, with the Soft Actor-Critic (SAC) being a prime candidate. SAC's utilization of artificial neural networks allows for a more effective generalization across the state-action space, following a model-free, off-policy framework. Notably, SAC not only supports our existing custom reward functions but is also adept at managing a continuous action and state space. Furthermore, implementing SAC paves the way for the inclusion of additional buildings in our simulations, specifically extending from 2 to all 17 buildings in the CityLearn dataset. This broader scope significantly enriches the training process and improves the accuracy of our data, culminating in more refined and dependable optimizations for the energy systems of grid-interactive efficient buildings for both cost and energy.

4. Contributions

In this project, each team member made significant contributions. Anna provided expertise on energy grid interactions with buildings and strategies for energy efficiency. Her responsibilities included data preprocessing, writing the background and environmental context sections of the report, and analyzing the project's limitations and future work. Bradley implemented the CityLearn environment as well as tabular Q-learning, and documented the results in the report. Jenna was responsible for designing the reward functions and writing the methodology section of the report. Together, we analyzed the results and produced the graphs and tables that illustrated the project's findings.

References

[1] U.S. Energy Information Administration (EIA), "Energy consumption: Residential, commercial, and industrial sectors," Nov 2023. [Online]. Available: https://www.eia.gov/totalenergy/data/monthly/pdf/sec2.pdf 1

[2] United States Environmental Protection Agency, "Sources of greenhouse gas emissions," Online Database, Nov 2023. [Online]. Available: https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions 1

[3] B. D. Leibowicz, C. M. Lanham, M. T. Brozynski, J. R. Vázquez-Canteli, N. C. Castejón, and Z. Nagy, "Optimal decarbonization pathways for urban residential building energy services," *Applied Energy*, vol. 230, pp. 1311–1325, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261918313552 1

[4] B. Jones-Albertus, "Confronting the duck curve: How to address over-generation of solar energy," Oct 2017. [Online]. Available: https://www.energy.gov/eere/articles/confronting-duck-curve-how-address-over-generation-solar-energy 1

[5] M. Neukomm, V. Nubbe, and R. Fares, "Grid-interactive efficient buildings technical report series: Overview of research challenges and gaps," U.S. Department of Energy Office of Energy Efficiency and Renewable Energy, Tech. Rep., December 2019. [Online]. Available: https://www1.eere.energy.gov/buildings/pdfs/75470.pdf 2

[6] C. Carmichael, M. Jungclaus, P. Keuhn, and K. P. Hydras, "Value potential for grid-interactive efficient buildings in the gsa portfolio: A cost-benefit analysis," 2019. [Online]. Available: https://rmi.org/insight/value-potential-for-grid-interactive-efficient-buildings-in-the-gsa-portfolio-a-cost-benefit-analysis/ 2

[7] J. R. Vázquez-Canteli, J. Kämpf, G. Henze, and Z. Nagy, "Citylearn v1.0: An openai gym environment for demand response with deep reinforcement learning," in *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ser. BuildSys '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 356–357. [Online]. Available: https://doi.org/10.1145/3360322.3360998 2

[8] K. Nweye, S. Sankaranarayanan, and Z. Nagy, "Merlin: Multi-agent offline and transfer learning for occupant-centric operation of grid-interactive communities," *Applied Energy*, vol. 346, p. 121323, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261923006876 3

[9] C. Yang, J. Zhang, F. Lin, L. Wang, W. Jiang, and H. Zhang, "Combining forecasting and multi-agent reinforcement learning techniques on power grid scheduling task," in *2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, 2023, pp. 1576–1580. 3

[10] "Los angeles intl ap 722950 (tmy3)," EnergyPlus, December 11 2023, https://energyplus.net/weather-location/north_and_central_america_wmo_region_4/USA/CA/USA_CA_Los.Angeles.Intl.AP.722950_TMY3. 3

[11] SCE.Com. Time-of-use residential rate plans. https://www.sce.com/residential/rates/Time-Of-Use-Residential-Rate-Plans. Accessed 11 Dec. 2023. 3