OfficeLearn: An OpenAI Gym Environment for Reinforcement Learning on Occupant-Level Building's Energy Demand Response

Lucas Spangher [1] **Akash Gokul** [1] **Joseph Palakapilly** [1] **Utkarsha Agwan** [1] **Manan Khattar** [1] **Wann-Jiun Ma** [2] **Costas Spanos** [1]

Abstract

Energy Demand Response (DR) will play a crucial role in balancing renewable energy generation with demand as grids decarbonize. There is growing interest in developing Reinforcement Learning (RL) techniques to optimize DR pricing, as pricing set by electric utilities often cannot take behavioral irrationality into account. However, so far, attempts to standardize RL efforts in this area do not exist. In this paper, we present a first of the kind OpenAI gym environment for testing DR with occupant level building dynamics. We demonstrate the flexibility with which a researcher can customize our simulated office environment through the explicit input parameters we have provided. We hope that this work enables future work in DR in buildings.

1. Introduction

Efforts to address climate change are impossible without a quick and safe transition to renewable energy. Barring strategies to address the "volatility" of generation in renewable energy sources, grids with increasing shares of renewable energy will face daunting consquences. These range from wasting of energy through curtailment (i.e., the shutting off of green energy sources when generation exceeds supply) (Spangher et al., 2020), voltage instability, or damage to the physical infrasture of the grid. Indeed, the California grid of 2019 needed to curtail roughly 3% of its energy, with some days seeing up to 33% of solar energy curtailed.

One solution to curtailment commonly touted is energy Demand Response (DR) which entails the deferment of energy demand from when it is demanded to when it is most opportune for it to be filled. DR is essentially costless, as it requires no infrastructure, so it is important as a direct solution.

One primary area of application for DR is in buildings. Buildings make up a significant, increasing component of US energy demand. In residential and commercial buildings, plug loads represent 30% of total electricity use ((Lanzisera et al., 2013), (Srinivasan et al., 2011)). In addition, the quantity of energy used by plugs is increasing more quickly than any other load type in both residential and commercial buildings (Comstock & Jarzomski, 2012).

Machine Learning (ML), while transformative in many sectors of the economy, is somewhat underdeveloped when it comes to energy applications. The creation of the AI in Climate Change community exist to bridge this gap, encouraging the collaboration of research to develop and apply a broad array of techniques into an equally broad array of applications. To encourage exploration in occupant level building DR, we propose to formalize an OpenAI Gym environment for the testing of Reinforcement Learning (RL) agents within a single office building.

2. Related Works

Deep RL is a subfield in ML that trains an agent to choose actions that maximize its rewards in an environment (Sutton & Barto, 1998). RL has had extensive success in complex control environments like Atari games (Mnih et al., 2013) and in previously unsolved domains like PSPACE-complete Sokoban planning (Feng et al., 2020). It has limited success in energy: Google implemented RL controls to reduce energy consumption in data centers by 40%. When DeepMind used a form of multi-agent RL to beat the world champion in the complex and strategic game of Go (Borowiec, 2016), researchers called for similar advancement in RL for power systems (Li & Du, 2018).

OpenAI Gym environments are a series of standardized environments that provide platform for benchmarking the progress of learning agents (Brockman et al., 2016). Gym environments allow researcher to create custom environments under a general and widely accepted API framework/format that immediately allows deployment of a suite of out-of-the-box RL techniques; therefore therefore so Gym environments tend to concentrate work around the specific problem that they describe.

Other groups have produced OpenAI Gym environments around similar goals but with different scopes. CityLearn aims to integrate multi-agent RL applications for DR in connected communities (Vázquez-Canteli et al., 2019). The environment supports customizing an array of variables in the creation of a heterogenous group of buildings, including number of buildings, type of buildings, and demand profile. A competition was hosted in CityLearn, in which the creators solicited submissions of agents that could learn appropriately in their environment (Kathirgamanathan et al., 2020).

We are unaware of an effort that attempts to focus study around occupant level energy DR in a Gym environment: that is, an effort that focuses on occupant level DR within a building. Therefore, we endeavor in this work to present the OfficeLearn Gym environment, an environment that may serve as preparation grounds for experiments implementing DR within real world test buildings.

2.1. Paper Outline

We have contextualized the creation of our Gym environment within the broader effort of applying ML techniques to climate change in Sections 1 and 2. In Section 3, we describe the technical details of our environment and how those will differ in future iterations. In Section 4, we illustrate the dynamics of the system by comparing key design choices in environmental setup. In Section 5, 6, and 7, we conclude, note how you might use the environment, and discuss future directions.

3. Description of Gym Environment

3.1. Overview

In this section, we highlight a summary of the environment and the underlying Markov Decision Process. The flow of information is succinctly expressed in Figure 1.

The environment takes the format of the following Markov Decision Process (MDP), (S, A, p, r):

- State Space S: The prices, energy usage, and baseline energy, all 10-dim vectors.
- Action Space A: A 10-dim vector (continuous or discrete) containing the agent's points.
- Transition probability p.
- Reward r defined in Section 3.5.

We describe out design choices and variants of the MDP below.

3.2. State space

The steps of the agent are currently formulated day by day, with ten-hour working days considered. Therefore, while the state space has several different components (described below), each is of ten dimensions as each one is hourly in nature.

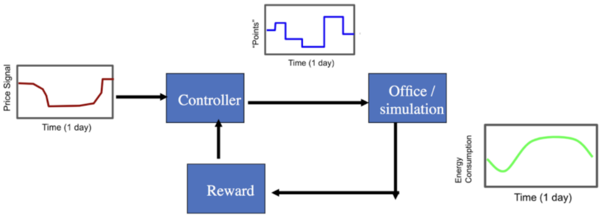

Figure 1. A schematic showing the interplay between agent and office environment, and ensuing energy reponses. The agent receives prices from the grid, then transforms it into "points" (called as such for differentiation.) Office workers engage with the points in the way an individual might be engaged with their home energy bill, which is reasonable assuming behavioral incentives detailed in (Spangher et al., 2020). The office recieves these points at the beginning of the "day". Workers proceed to use energy throughout the day and at night the system delivers a record of their energy consumption, which is reduced into a reward that trains the agent.

3.2.1. Grid price regimes

Utilities are increasingly moving towards time dependent energy pricing, especially for bigger consumers such as commercial office buildings with the capacity to shift their energy usage. Time of use (TOU) pricing involves is a simple, two-level daily price curve that changes seasonally and is declared ahead of time. We use PG&E's TOU price curves from 2019. Real time pricing (RTP), meanwhile, is dynamic for every hour and changes according to supply and demand in the energy market. We simulate it by subtracting the solar energy from demand of a sample building. There is significant seasonal variation in prices depending on geography, e.g. in warmer climates, the increased cooling load during summer can cause an increase in energy prices.

3.2.2. Energy of the prior steps

The default instantiation of the environment includes the energy use of office workers of the prior step. This allows the agent to directly consider a day-to-day time dependence. The simulated office workers in this version are currently memoryless day to day in their energy consumption, but a future simulation will allow for weekly deferable energy demands to simulate weekly work that can be deferred and then accomplished.

The energy of the prior steps may be optionally excluded from the state space by those who use our environment.

3.2.3. Grid prices of the prior step

Users may optionally include the grid price from prior steps in the state space. This would allow the agent to directly consider the behavioral hysteresis that past grid prices may have on a real office worker's energy consumption. Although this is a noted phenomenon in human psychology generally (Richards & Green, 2003), it is not well quantified and so we have not included it in how we calculate our simulated human agents.

3.2.4. Baseline energy

Baseline Energy may optionally be included in the state space. If the agent directly observes its own action and the baseline energy, it observes all of the information necessary to calculate certain simpler simulated office worker responses. Therefore, inclusion of this element will make the problem fully observable, and truly an MDP rather than Partially Observable MDP (POMDP).

3.3. Action space

The agent's action space expresses the points that the agent delivers to the office. The action space is by default a continuous value between zero and ten, but may be optionally discretized to integer values if the learning algorithm outputs discrete values.

The purpose of the action is to translate the grid price into one that optimizes for behavioral response to points. Therefore, the policy will learn over time how people respond to the points given and maximally shift their demand towards the prices that the grid gives.

3.4. Office workers: simulated response functions

In this section, we will summarize various simulated responses that office workers may exhibit.

3.4.1. "Deterministic Office Worker"

We include three types of deterministic response, with the option for the user to specify a mixed office of all three.

In the linear response, we define simple office worker who decreases their energy consumption linearly below a baseline with respect to points given. Therefore, if b_t is the baseline energy consumption at time t and p_t are the points given, the energy demand d at time t is $d_t = b_t - p_t$, clipped at d_{min} and d_{max} as defined in Section 3.5.

In the sinusoidal response, we define an office worker who responds well to points towards the middle of the distribution and not well to prices at the. Therefore, the energy demand d at time t is $d_t = b_t - \sin p_t$, clipped at d_{min} and d_{max}.

In the threshold exponential response, we define an office worker who does not respond to points until they are high, at which point they respond exponentially. Therefore, the energy demand d is $d_t = b_t - (\exp p_t * (p_t > 5))$, clipped at d_{min} and d_{max}.

3.4.2. "Curtail And Shift Office Worker"

Office workers need to consume electricity to do their work, and may not be able to curtail their load below a minimum threshold, e.g. the minimum power needed to run a PC. They may have the ability to shift their load over a definite time interval, e.g. choosing to charge their laptops ahead of time or at a later time. We model a response function that exhibits both of these behaviors. We can model the aggregate load of a person (b_t) as a combination of fixed inflexible demand (b_t^{fixed}), curtailable demand ($b_t^{curtail}$), and shiftable demand (b_t^{shift}), i.e., $b_t = b_t^{fixed} + b_t^{curtail} + b_t^{shift}$. All of the curtailable demand is curtailed for the $T_{curtail}$ hours (set to 3 hours in practice) with the highest points, and for every hour t the shiftable demand is shifted to the hour within $[t - T_{shift}, t + T_{shift}]$ with the lowest energy price.

3.5. Reward

Specification of the reward function is notoriously difficult, as it is generally hand-tailored and must reduce a rich and often multi-dimensional environmental response into a single metric. Although we include many possible rewards in the code, we outline the two rewards that we feel most accurately describe the environment. As we already demonstrated in prior work the ability to reduce overall energy consumption (Spangher et al., 2019), we endeavor to direct this agent away from reducing consumption and towards optimally shifting energy consumption to favorable times of day.

3.5.1. Scaled Cost Distance

This reward is defined as the difference between the day's total cost of energy and the ideal cost of energy. The ideal cost of energy is obtained using a simple convex optimization. If $\vec{d}$ are the actual demand of energy computed for the day, $\vec{g}$ is the vector of the grid prices for the day, E is the total amount of energy, and d_{min}, d_{max} are 5% and 95% values of energy observed over the past year, then the ideal demands are calculated by optimizing the objective: $d^* = \min_d d^T g$ subject to the constraints $\sum_{t=0}^{10} d = E$ and $d_{min} < d < d_{max}$. Then, the reward becomes: $R(d) = \frac{d^{*T} g - d^T g}{d^{*T} g}$, i.e. taking the difference and scaling by the total ideal cost to normalize the outcome.

3.5.2. Log Cost Regularized

Although the concept of the ideal cost is intuitive, the simplicity of the convex optimizer means that the output energy is often an unrealistic, quasi step function. Therefore, we propose an alternate reward of log cost regularized. Following the notation from above, the reward is $R(d) = -d^T g - \lambda(\sum d < 10 * (.5 * b_{max}))$, where b_{max} refers to the max value from the baseline. In practice, we set λ to some high value like 100. The purpose of the regularizer is to penalize the agent for driving down energy across the domain, and instead encourage it to shift energy.

4. Illustration of Features

We will now demo the environment's functioning. All comparisons are done with a vanilla Soft Actor Critic RL agent that learns throughout 10000 steps (where one step is equal to one day), with a TOU pricing regime fixed at a single day. The agent's points are scaled between -1 and 1.

4.0.1. Comparison of Rewards type

We present the effect of using the Log Distance Regularized and the Scaled Cost Distance. Please see Figure 2, in the Appendix, for side by side comparison of the reward types. In this figure, you can see that not only is the agent capable of learning an action sequence that accomplishes a lower cost than if the simulated office workers were to respond directly to the untransformed grid prices, but also differs in how the learning is guided. The log cost regularized reward accomplishes smoother prices that result in the agent deferring most of the energy for the end of the day, whereas the scaled cost distance reward allows for more energy earlier in the day, guiding the simulated office worker to increase energy gradually throughout the day.

4.0.2. Comparison of office worker response functions

We present the effect of using different simulated office workers on the output of energy demand. Please see Figure 3, in the Appendix, for a comparison of two types of simulated office workers. In the exponential response, we see an example of how the office worker's energy demand responds to points – that is, perhaps, too coarsely for a learner to make much difference. Meanwhile, the Curtail and Shift response demonstrates a much richer response, which enables a learner to learn the situation and perform better than the control.

5. Conclusion

We present technical details of a novel gym environment for the testing of RL for energy DR within a single building. We detail the design choices that we made while constructing the environment. We then show demos of different reward types and simulated office responses.

6. Simulating DR in your building

The environment we provide contains many ways to customize your own building. You may choose the number of occupants, their response types, baseline energies, grid price regimes, and frequency with which grid price regimes change. You may also choose from a host of options when it comes to customizing the agent and its state space. Please contact us if you are interested in deeper customization and would like a tutorial on the code.

7. Future Work

7.1. Variants of the MDP

We plan to offer the user the choice between a step size that is a day's length and a step size that is an hour's length. The alteration can provide a more efficient state space representation that provides for a fully observable MDP for the agent, as well as a longer trajectory for action sequences (i.e., ten steps for every trajectory to determine the ten hours rather than a single step producing all ten hours), at which RL tends to excel.

7.2. Reality Gap

Similar to existing simulations, e.g. Sim2Real (CITE SIM2REAL), there is a gap between our environment and reality. Future work in this direction will build more realistic response functions by relying on existing modelling literature (CITE Planning Circuit from Fall'19).

7.3. OfficeLearn Competition

We plan to host a OfficeLearn competition in the future. This competition will prioritize agents that can maximize sample efficiency, due to the realistic time constraints of a social game, and deferability of energy on a test set of simulated office workers.

8. Acknowledgements

We would like to thank our wonderful and responsive peers Peter Henderson and Andreea Bobu for their thoughtful advice and commentary on the environment. We would like to also extend thanks to Alex Devonport, Adam Bouyamourn, and Akaash Tawade for their help in earlier versions. We would all, finally, like to thank our respective parents.

References

Borowiec, S. Alphago seals 4-1 victory over go grandmaster lee sedol. *The Guardian*, 15, 2016.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym, 2016.

Comstock, O. and Jarzomski, K. Consumption and saturation trends of residential miscellaneous end-use loads. *ACEEE Summer Study on Energy Efficiency in Buildings, Pacific Grove, CA, USA*, 2012.

Feng, D., Gomes, C. P., and Selman, B. Solving hard ai planning instances using curriculum-driven deep reinforcement learning. *arXiv preprint arXiv:2006.02689*, 2020.

Kathirgamanathan, A., Twardowski, K., Mangina, E., and Finn, D. A centralised soft actor critic deep reinforcement learning approach to district demand side management through citylearn, 2020.

Lanzisera, S., Dawson-Haggerty, S., Cheung, H. Y. I., Taneja, J., Culler, D., and Brown, R. Methods for detailed energy data collection of miscellaneous and electronic loads in a commercial office building. *Building and Environment*, 65:170–177, 2013.

Li, F. and Du, Y. From alphago to power system ai: What engineers can learn from solving the most complex board game. *IEEE Power and Energy Magazine*, 16(2):76–84, 2018.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Richards, T. J. and Green, G. P. Economic hysteresis in variety selection. *Journal of Agricultural and Applied Economics*, 35(1):1–14, 2003.

Spangher, L., Tawade, A., Devonport, A., and Spanos, C. Engineering vs. ambient type visualizations: Quantifying effects of different data visualizations on energy consumption. In *Proceedings of the 1st ACM International Workshop on Urban Building Energy Sensing, Controls, Big Data Analysis, and Visualization*, UrbSys'19, pp. 14–22, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450370141. doi: 10.1145/3363459.3363527. URL https://doi.org/10.1145/3363459.3363527.

Spangher, L., Gokul, A., Khattar, M., Palakapilly, J., Tawade, A., Bouyamourn, A., Devonport, A., and Spanos, C. Prospective experiment for reinforcement learning on demand response in a social game framework. In *Proceedings of the 2nd International Workshop on Applied Machine Learning for Intelligent Energy Systems (AMLIES) 2020*, 2020.

Srinivasan, R. S., Lakshmanan, J., Santosa, E., and Srivastav, D. Plug load densities for energy analysis: K-12 schools,. *Energy and Buildings*, 43:3289 – 3294, 2011.

Sutton, R. S. and Barto, A. G. Reinforcement learning i: Introduction, 1998.

Vázquez-Canteli, J. R., Kämpf, J., Henze, G., and Nagy, Z. 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*, BuildSys '19, pp. 356–357, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450370059. doi: 10.1145/3360322.3360998. URL https://doi.org/10.1145/3360322.3360998.

9. Appendix

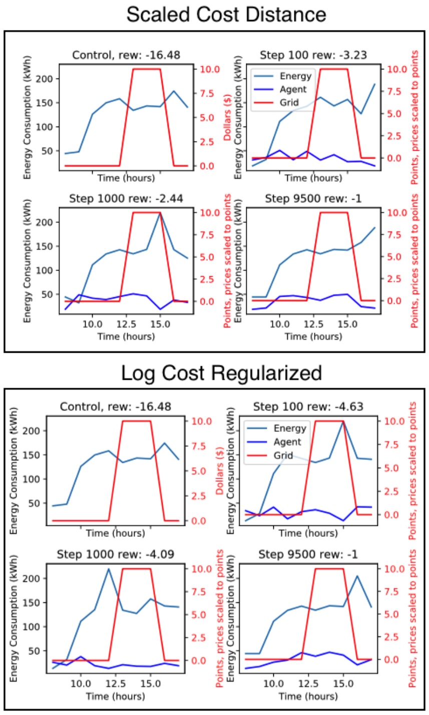

Figure 2. A comparison of the Log Cost Regularized and the Scaled Cost Distance rewards. The energy output of the simulated office workers is drawn in light blue, and corresponds to the primary axes. The grid prices are drawn in red, and refers to TOU pricing. It corresponds to the secondary axes. The agent's actions are drawn in dark blue, is scaled between -1 and 1 to improve readability of the plots, and correspond to the secondary axes.

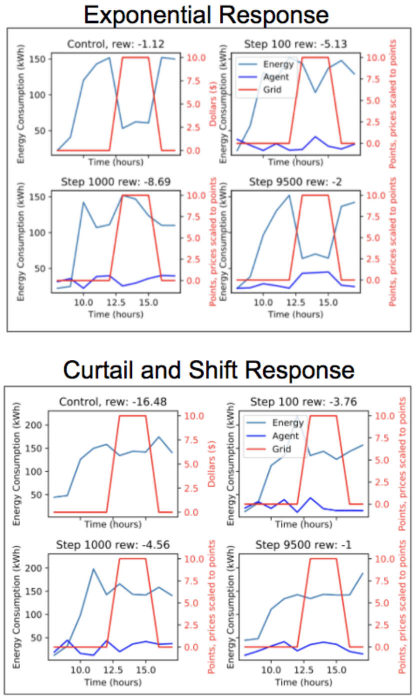

Figure 3. A comparison of the "Exponential Deterministic Office Worker" to the "Curtail and Shift Office Worker". The energy output of the simulated office workers is drawn in light blue, and corresponds to the primary axes. The grid prices are drawn in red and corresponds to the secondary axes. The agent's actions are drawn in dark blue, is scaled between -1 and 1, and correspond to the secondary axes.