Deep Reinforcement Learning for Energy Management in Residential Housing

Tim Walter [1]

Abstract

Buildings account for a significant portion of global energy consumption and emissions. This paper explores how smart energy management systems in single-family homes can potentially reduce emissions, focusing on homes equipped with photovoltaic systems, electric batteries, and system-controllable appliances. The aim is to optimize energy consumption and generation capacities to minimize emissions, manage load flexibility, and alleviate grid pressure from fluctuating renewable energy sources. The controller also enables the sale of cleanly generated electricity to the grid at a discounted emission premium. The control mechanisms involve deep reinforcement learning algorithms benchmarked against traditional thresholding models and the theoretical optimum.

1. Introduction

Introduction about climate change and how our power generation has to adapt and how we can help to adapt our demand to the newly fluctuating energy generation

2. Related Work

The application field of smart energy management encompasses a wide range of problems, such as heating and cooling (Blum et al., 2021)(Thomas Schreiber et al., 2020), flexible demand response (Jin et al., 2021) and energy storage (Nakabi & Toivanen, 2021). Moreover, the scale and level of detail of the tackled problems vary greatly, going from single buildings (Blum et al., 2021), to microgrids (Nakabi & Toivanen, 2021) (Castellanos et al.) to eventually continent wide electricity grids (Hörsch et al., 2018). The two most promising approaches to solve such problems is the classical Model Predictive Control (MPC) (Basantes et al., 2023) and Reinforcement Learning (RL) (Nakabi & Toivanen, 2021) (Jin et al., 2021) (Thomas Schreiber et al., 2020)(Zhu et al., 2022). Furthermore, there are also hybrid approaches that combine the two methods (Javier Arroyo et al., 2022). While MPC is a well established control method, RL is a relatively new approach that has gained a lot of attention in the last years. The main advantage of RL is that it does not require a model of the environment, which is often hard to obtain in the energy domain. However, RL is known to be sample inefficient and requires a lot of training data.

This paper is mainly inspired by (Nakabi & Toivanen, 2021), who describe a microgrid environment with generation, storage and consumers and optimized electricity pricing using various deep reinforcement learning algorithms.

3. Environment

The key contribution of this work is to build an environment for a single family household, consisting of a rooftop solar array for energy generation, an energy storage system, a thermostatically controlled load and flexible demand response. The enviroment is built on a mixture of simulated and real data and dynamic components. It is used to optimize the CO2eq emissions of the household. The environment is built on the Gymansium framework (Towers et al., 2023) and is available on GitHub[1]

The training environment was chosen to be episodic to accomodate the environment framework(Towers et al., 2023) and algorithm framework(Antonin Raffin et al., 2021). The episodes are organized into weeks, with either minutely or hourly resolution, to capture the daily and weekly patterns of the household. The timeframe chosen, was 07.01.2007 - 28.12.2008, although some episodes had to be left out due to missing data. Overall there were over 100 episodes, of which 95 were used for training, 4 for testing and the rest for training. The power was standarized to kilowatts, while the energy was simply kWmin or kWh depending on the resolution. As the agents is operating on equidistant timesteps anyway, there is no real difference between energy and power, as the power is always assumed to be constant

[1]Department of Scientifc Computing, Technical University of Munich, Munich, Germany. Correspondence to: Tim Walter <tim.walter@tum.de>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]https://github.com/TimWalter/smart-energy-controller

over the timestep. The location of the house was in France near Paris.

3.1. Static Components

The static components are the parts of the environment that are not reactive to the agent's actions. These are auxillary information, the generation of the rooftop solar array, the household energy demand and the external electricity supply. The data was either given in hourly or minutely resolution, and was adapted to the missing resolution using linear interpolation or averaging.

3.1.1. Auxillary Information

As auxillary information, weather and time data was given, which was sourced from the Photovoltaic Geographical Information System (PVGIS) (Thomas Huld et al., 2012). They had no direct influence on the reward and only provided means for the agent to predict the carbon intensity of the electricity mix. The only exception was the outdoor temperature, which influenced the thermostatically controlled load. The actual observables of this component are given in table 1.

3.1.2. Rooftop Solar Array (RSA)

The generated energy was simulated using the PVLIB library (F. Holmgren et al., 2018). The simulation used the weather data described in 3.1.1 and the solar panel specifications from the CEC database (Dobos, 2012)(Boyson et al., 2007).

3.1.3. Household Energy Demand

The household energy demand was sourced from the UCI archive (Georges Hebrail). Since the consumption was unraveled for different rooms, the kitchen, electric water heater and air conditioner were modelled as inflexible demand, while the laundry room was modelled as flexible demand.

3.1.4. External Electricity Supply (EES)

The direct carbon intensity of the electricity mix at any given time was sourced from electricity maps (.29, 29.12.2023) for France. However, only data from 2021 and 2022 was available, so the carbon intensity might be biased, since the electricity mix might have drifted over the years. However, the daily, weekly and yearly frequencies should still be captured, although potentially drifted. The unit of carbon intensity is gramm of CO_2 equivalent emissions per kWh or kWmin.

To showcase the scenario and patterns the static components are visualized for the first 24 hours in figure 1, excluding the auxillary information.

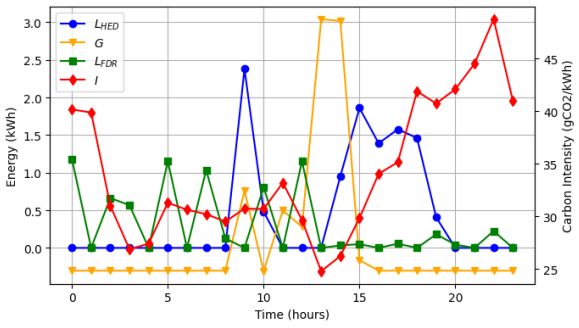

Figure 1. Static components for the first 24 hours

3.2. Dynamic Components

The dynamic components consist of an energy storage system, flexible demand response and thermostatically controlled load.

3.2.1. Energy Storage System

The energy storage system (ESS) is directly chargable and dischargeable from either the EES or the RSA. The charge is subject to the followng dynamics:

$$B_t = \max\{0, B_{t-1} - D_s + C_t\sqrt{\nu} - \frac{D_t}{\sqrt{\nu}}\} \quad (1)$$

Here, $B_t \in [0, B_{max}]$ is the charge at time t and B_{max} is the capacity. Furthermore, ν denotes the round trip efficiency, D_s the self-discharge rate, $C_t \in [0, C_{max}]$ the charge rate with maximum C_{max} and $D_t \in [0, D_{max}]$ the discharge rate with its maximum D_{max}. The rates can be further constrained by the current charge and capacity, such that actually $C_t \in [0, \min\{C_{max}, \frac{B_{max}-B_t+D_s}{\sqrt{\nu}}\}]$ and $D_t \in [0, \min\{D_{max}, (B_t - D_S)\sqrt{\nu}\}]$.

3.2.2. Flexible Demand Response

Flexible Demand Response (FDR) denotes the shifting of load in time. This is modelled as a desired schedule of usage, as described in 3.1.3. A timeframe centred around the current timestep of the schedule is influenced stochatistacally by a scalar signal $a_{fdr,t} \in [-1, 1]$. Whether a usage is executed is determined by a Bernoulli process, with probabilities p calculated from the desired schedule, the signal and an exponential weighting, which is centered around the scheduled time of usage. The probabilities are given by:

$$p_t = \max\{0, \min\{1, s + a_{fdr,t} * \exp(\frac{-1}{\beta}|t - t_s|)\}\} \quad (2)$$

where $s \in \{0, 1\}^H$ is the vector of the desired schedule with elements $s_i = \begin{cases} 1 & \text{if } t_i \leq t \\ 0 & \text{else} \end{cases}$ and H the number of

usages in the timeframe, β is a patience parameter, and $t_s \in \mathbb{R}^H$ is the vector of desired execution times in the current timeframe. If a usage is executed its power is completely drawn in the current timestep. However, if a usage is not executed after being delayed once, it will consume a discounted amount of power in each timestep, such that if it is never exectued it will consume the same amount of power as if it was executed once. This aims to facilitate learning by preventing infinite delays or a large negative reward at the end of the episode. To do so the power consumption of a usage after the first delay is $\frac{2p}{H-1}$, which is also discounted from the remaining power.

3.2.3. Thermostatically Controlled Load

The Thermostatically Controlled Load (TCL) encapsulates all devices that aim to maintain the temperature of a given heat mass. To utilize such loads as energy storage, the temperature of the heat mass is allowed to fluctuate within a given range. The TCL is modeled as a second-order system (Sonderegger, 1978) that has the following update procedure:

$$\begin{aligned} T_t = T_{t-1} &+ \frac{1}{r_a}(T_{a,t} - T_{t-1}) \\ &+ \frac{1}{r_b}(T_{b,t} - T_{t-1}) \\ &+ \frac{1}{r_h} L_{TCL} a_{tcl,t} + q \end{aligned} \tag{3}$$

where T is the indoor temperature, r_a is the thermal mass of the air, T_a is the outdoor temperature, r_b is the thermal mass of the building material, and T_b is the building mass temperature, that evolves with

$$T_{b,t} = T_{b,t-1} + \frac{1}{r_b}(T_{t-1} - T_{b,t-1}) \tag{4}$$

, $\frac{1}{r_h}$ is the power-to-heat coefficient, L_{TCL} is the nominal power, $a_{tcl_t} \in [-1, 1]$ is the heating signal, and q is the unintentional heat drift. The heating signal is constrained by the desired temperature range, enforced through a backup controller as follows:

$$a_{tcl,t} = \begin{cases} -1 & \text{if } T_t \geq T_{max} \\ 1 & \text{if } T_t \leq T_{min} \\ a_{tcl,t} & \text{else} \end{cases} \tag{5}$$

where T_{max} and T_{min} define the desired temperature range.

3.3. Markov Decision Process

The data in combination with the dynamic components allows the formulation of a Markov Decision Process (MDP), necessary for the RL paradigm. This requires the definition of an observation space, action space, a reward function and a transition function. The transition function is given by the replay of measurements described in 3.1 and the dynamics described in 3.2.

3.3.1. Observation Space

The observation space is a bounded subset of $\mathbb{R}^{12}$, with the observables given in 1.

Observable	**Symbol**
Carbon Intensity	I
Household Energy Demand	L_{HED}
Rooftop Solar Generation	G
ESS Charge	B
FDR power in current timeframe	L_{FDR}
Indoor temperature	T
Day of Year	DoY
Hour of Day	HoD
Solar Irradiation	SI
Solar Elevation	SE
Outdor temperature	T_a
Wind Speed	WS

Table 1. Observation Space

3.3.2. Action Space

The agent can influence the components of 3.2, through the signals a_{fdr}, a_{tcl} and a_{ess}. For a_{ess} the charging and discharging rates are combined into a single signal, since the ESS can either be charged or discharged at any given timestep. All signal are normalized to $[-1, 1]$ and can be discretized into bins, if the algorithm require a discrete action space. The action space is summarized in table 2.

Action	**Min. Meaning**	**Max. Meaning**
a_{ess}	Discharge	Charge
a_{fdr}	Delay	Expedite
a_{tcl}	No heating	Heating

Table 2. Action Space

3.3.3. Reward Function

The reward function is mainly given by the physical quantity of CO2eq emissions. However, to prevent unintended exploitation of the TCL, which is often optimal at constant T_{min}, it is augmented by a discomfort penalty. The resulting

function is given by:

$$r_t = I_t(PE - CE) - DC \tag{6}$$

$$EC = L_t + \frac{1}{r_h} L_{TCL} a_{tcl,t} + C_t + \sum_{p \in U_e} p_r + \sum_{p \in U_d} \frac{2p}{H-1} \tag{7}$$

$$E_p = G_t + D_t \tag{8}$$

$$DC = \delta exp(|T_t - \frac{T_{max} + T_{min}}{2}) \tag{9}$$

where PE is the energy produced, CE is the energy produced, U_e is the set of executed usages, p_r is the remaining power after discounting, U_d is the set of delayed usages, DC is the discomfort with a penalty factor δ.

Moreover, since the task is episodically formulated, there is a correction necessary at the end of the episode to account for e.g. differences in charge left. The correction of the final reward is given as:

$$reward \mathrel{+}= I_t * B_t \tag{10}$$

$$reward \mathrel{+}= I_t * (-\sum_{p \in U} p_r) \tag{11}$$

$$reward \mathrel{+}= I_t * (-r_h |T_t - \frac{T_{max} + T_{min}}{2}|) \tag{12}$$

where U is the set of left over usages. Every equation is correcting for one type of left over energy in the system.

4. Limitations

5. Algorithms

This section will describe the algorithms used to solve the MDP. The algorithms will be benchmarked against a theoretical optimum and a thresholding model, which is also described here.

5.1. Thresholding Model

A quick subsection about the non machine learning baseline

5.2. Theoretical Optimum

A quick subsection about the theoretical optimum

5.3. Reinforcement Learning

Subsection that presents the reinforcement learning algorithms used to solve the MDP

6. Results

This section will present the results of the algorithms and compare them to the baseline and theoretical optimum.

7. Discussion

References

Electricity maps — reduce carbon emissions with actionable electricity data, 29.12.2023. URL https://www.electricitymaps.com/.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Basantes, J. A., Paredes, D. E., Llanos, J. R., Ortiz, D. E., and Burgos, C. D. Energy management system (ems) based on model predictive control (mpc) for an isolated dc microgrid. *Energies*, 16(6):2912, 2023. doi: 10.3390/en16062912.

Blum, D., Arroyo, J., Huang, S., Drgoňa, J., Jorissen, F., Walnum, H. T., Chen, Y., Benne, K., Vrabie, D., Wetter, M., and Helsen, L. Building optimization testing framework (boptest) for simulation-based benchmarking of control strategies in buildings. *Journal of Building Performance Simulation*, 14(5):586–610, 2021. ISSN 1940-1493. doi: 10.1080/19401493.2021.1986574.

Boyson, W. E., Galbraith, G. M., King, D. L., and Gonzalez, S. Performance model for grid-connected photovoltaic inverters. 2007. doi: 10.2172/920449. URL https://www.osti.gov/biblio/920449,journal=.

Castellanos, J., Correa-Florez, C. A., Garcés, A., Ordóñez-Plata, G., Uribe, C. A., and Patino, D. An energy management system model with power quality constraints for unbalanced multi-microgrids interacting in a local energy market. URL http://arxiv.org/pdf/2212.01910.pdf.

Dobos, A. P. An improved coefficient calculator for the california energy commission 6 parameter photovoltaic module model. *Journal of Solar Energy Engineering*, 134 (2), 2012. ISSN 0199-6231. doi: 10.1115/1.4005759.

F. Holmgren, W., W. Hansen, C., and A. Mikofski, M. pvlib python: a python package for modeling solar energy systems. *Journal of Open Source Software*, 3(29):884, 2018. doi: 10.21105/joss.00884.

Georges Hebrail, A. B. Individual household electric power consumption.

Hörsch, J., Hofmann, F., Schlachtberger, D., and Brown, T. Pypsa-eur: An open optimisation model of the european transmission system. *Energy Strategy Reviews*, 22:207–215, 2018. ISSN 2211467X. doi: 10.1016/j.

esr.2018.08.012. URL http://arxiv.org/pdf/1806.01613.pdf.

Javier Arroyo, Carlo Manna, Fred Spiessens, and Lieve Helsen. Reinforced model predictive control (rl-mpc) for building energy management. *Applied Energy*, 309:118346, 2022. ISSN 0306-2619. doi: 10.1016/j.apenergy.2021.118346. URL https://www.sciencedirect.com/science/article/pii/S0306261921015932.

Jin, L., Chen, Z., Li, J., and Ye, T. Reinforcement learning method based load shifting strategy with demand response. In *2021 40th Chinese Control Conference (CCC)*, pp. 1586–1591, 2021. doi: 10.23919/CCC52363.2021.9549975.

Nakabi, T. A. and Toivanen, P. Deep reinforcement learning for energy management in a microgrid with flexible demand. *Sustainable Energy, Grids and Networks*, 25:100413, 2021. ISSN 23524677. doi: 10.1016/j.segan.2020.100413.

Sonderegger, R. C. *Dynamic models of house heating based on equivalent thermal parameters*. PhD thesis, Princeton University, New Jersey, 1978.

Thomas Huld, Richard Müller, and Attilio Gambardella. A new solar radiation database for estimating pv performance in europe and africa. *Solar Energy*, 86(6):1803–1815, 2012. ISSN 0038-092X. doi: 10.1016/j.solener.2012.03.006. URL https://www.sciencedirect.com/science/article/pii/S0038092X12001119.

Thomas Schreiber, Sören Eschweiler, Marc Baranski, and Dirk Müller. Application of two promising reinforcement learning algorithms for load shifting in a cooling supply system. *Energy and Buildings*, 229:110490, 2020. ISSN 0378-7788. doi: 10.1016/j.enbuild.2020.110490. URL https://www.sciencedirect.com/science/article/pii/S0378778820320922.

Towers, M., Terry, J. K., Kwiatkowski, A., Balis, J. U., de Cola, G., Deleu, T., Goulão, M., Kallinteris, A., KG, Arjun, Krimmel, M., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J. J., Shen, A. T. J., and Younis, O. G. Gymnasium, 2023. URL https://zenodo.org/record/8127025.

Zhu, D., Yang, B., Liu, Y., Wang, Z., Ma, K., and Guan, X. Energy management based on multi-agent deep reinforcement learning for a multi-energy industrial park. *Applied Energy*, 311:118636, 2022. ISSN 0306-2619. doi: 10.1016/j.apenergy.2022.118636. URL http://arxiv.org/pdf/2202.03771.pdf.

A. You *can* have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one, even using the one-column format.