Outline Smart Energy Controller for Seminar on Machine learning for sequential decision making

Anonymous Authors[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. Environment

The environment and reward function formulate the problem for the agent to solve by finding an optimal policy. The key components are described in the following.

2.1. Photovoltaic System

Simulated using the PVLIB library, providing power per timestamp (G_t) as part of the state.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2.2. Battery

The primary energy storage with charge defined by:

$$B_t = B_{t-1} D_s + C_t \sqrt{\nu} - \frac{D_t}{\sqrt{\nu}} \tag{1}$$

Here, $B_t \in [0, B_{max}]$ is the charge at time t and B_{max} is its upper limit, these two values provide the normalized state of charge $BSoC_t = \frac{B_t}{B_{max}}$. Furthermore, ν denotes the round trip efficiency of the battery, D_s is the self-discharge rate, $C_t \in [0, C_{max}]$ is the charging rate and $D_t \in [0, D_{max}]$ is the discharging rate. The charging and discharging rates are combined into a single action for the agent, since the battery can either be charged or discharged at any given timestep.

$$a_{b,t} = \begin{cases} C_t, D_t = 0 & \text{if } a_{b,t} > 0 \\ D_t, C_t = 0 & \text{else} \end{cases} \tag{2}$$

where $a_{b,t} \in [B_t D_S \sqrt{\nu}, \frac{B_{max} - B_t D_s}{\sqrt{\nu}}]$.

2.3. Thermostatically Controlled Load

Under the term thermostatically controlled load (TCL) are all devices subsummed that aim to maintain the temperature of a given heat mass, such as a refrigerator, water boiler, or heat pump. 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

$$T_t = T_{t-1} + \frac{1}{c_a}(T_{out,t} - T_{t-1}) + \frac{1}{c_m}(T_{mass,t} - T_{t-1}) + L_{TCL} a_{tcl,t} + q \tag{3}$$

Where T_t is the measured indoor temperature, $T_{out,t}$ is the outdoor temperature, and $T_{mass,t}$ is the unobservable building mass temperature that evolves with

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

Furthermore, c_a and c_m are the thermal masses of the building and the air. Moreover, q denotes the unintentional heating of the building, L_{TCL} the nominal power of the TCL and $a_{tcl,t} \in [0, 1]$ the control decision, which is constrained by the desired temperature range, enforced through a backup

controller as follows:

$$a_{tcl,t} = \begin{cases} 0 & \text{if } T_t \geq T_{max} \\ 1 & \text{if } T_t \leq T_{min} \\ a_{tcl,t} & \text{else} \end{cases} \tag{5}$$

where T_{max} and T_{min} are the maximum and minimum temperature of the desired temperature range. Another normalized state variable is the state of charge of the indoor temperature $TSoC_t = \frac{T_t - T_{min}}{T_{max} - T_{min}}$.

2.4. Controllable Appliances

The usage of controllable appliances can be delayed or expendited by the agent. Such appliances could be e.g. a washing machine, dishwasher, or electric vehicle. Here, the inhabitants' behaviour is modelled stochastically. The agent has access to a preplanned schedule of usage for a given timeframe, and can decide to delay or expedite certain usages by defining in each timestep whether the appliance should be used or not. Whether it is actually used is then decided by a Bernoulli process with a probability of

$$p_t^i = e^{-\frac{1}{\beta}|t - t_s|} \tag{6}$$

with the probability mass function

$$f(a_{a,i,t}, p_t^i) = \begin{cases} p & \text{if } a_{a,i,t} = 1 \\ 1 - p & \text{if } a_{a,i,t} = 0 \end{cases} \tag{7}$$

where p_t^i is the probability of executing the action i at time t, t_s is the scheduled time of usage, and β is a patience parameter. The timeframe visible to the agent will be centered around the current timestep. If an action is not executed, while being in the visible timeframe, it is deterministically executed before exiting the timeframe. Contrary, if an action was executed and is still within the timeframe its power consumption is set to zero.

2.5. Uncontrollable Appliances

Uncontrollable appliances operate with a given power consumption and cannot be managed by the agent. They essentially contribute a negative generation L_t at each time step. This data isn't simulated; instead, it draws from real measurements taken from a French household's consumption tracked over four years.

2.6. Electricity Grid

The electricity grid serves as the primary source of reward for the agent. Selling power surpluses to the grid earns an emission premium, yielding positive rewards. Conversely, if the battery or generated power falls short of meeting the load, the grid automatically balances the deficit, resulting in a negative reward tied to the current CO2eq intensity and the energy consumed.

3. MDP Formulation

With this environment established, a Markov Decision Process (MDP) can be defined. The problem is episodic, organized into week-long episodes, and manages controlled appliance usage within daily timeframes. Initially, the timesteps are set to one minute but might be adjusted later. There are over one hundred episodes available, spanning a two-year period.

3.1. State

The state can be split into three categories: the controllable state S_c, the uncontrollable state S_u, and the informal state S_i. The controllable state describes all parts of the environment that the agent controls either directly or indirectly. These are the state of charge for the TCL and battery and a vector $s_{a,t} \in \mathbb{R}^{N \times 2}$, that describes the scheduled usage of the controllable appliances, in terms of power consumption and desired execution time, here N is the number of elements in the visible timeframe.

$$s_{c,t} = [BSoC_t, TSoC_t, s_{a,t}] \in [[0,1],[0,1],\{0,1\}^N] = S_c \tag{8}$$

The uncontrollable state describes all reward influencing parts of the environment that the agent cannot control. These are the current generation and load, as well as the CO2eq intensity for the upcoming timestep.

$$s_{u,t} = [G_t, L_t, I_t] \in [\mathbb{R}^+, \mathbb{R}^+, \mathbb{R}^+] = S_u \tag{9}$$

Lastly, the informal state describes all parts of the environment that are not directly influencing the reward but can be helpful for reward and state transition modelling. These are the current datetime t and the current weather W_t.

$$s_{i,t} = [t, W_t] \in S_i \tag{10}$$

The state is then given by the concatenation of the three state categories.

$$s_t \in S = S_c \times S_u \times S_i \tag{11}$$

3.2. Action

The agent can influence the enviroment according to the mechanisms described in the earlier section. Consequently, the action space is given by the actions on the battery $A_b = [0,1]$, on the TCL $A_{TCL} = [0,1]$, the selling of generated clean energy to the grid $A_s = [0, G_t]$, and scheduled appliances within the current timeframe $A_a = \{0,1\}^N$. The action space is then given by the concatenation of the three action spaces.:

$$a_t \in A = A_b \times A_{TCL} \times A_s \times A_a \tag{12}$$

3.3. Reward

The reward function might be subject to change but initially will punish the agent mainly for emitting CO2eq and reward it for selling clean energy to the grid. The reward function is given by

$$r_t = c_p a_{s,t} - c_e I_t (E_c - E_p) \tag{13}$$

$$E_c = L_t + a_{s,t} + a_{tcl,t} L_{TCL} + C_t + \sum_{i \in E_a}^{N} (s_{a,t})_{i,1} \tag{14}$$

$$E_p = G_t + D_t \tag{15}$$

where c_p is a coefficient determining the CO2eq premium, c_e is a coefficient determining the CO2eq emission penalty, E_c is the total energy consumed, E_p is the total energy produced, E_a is the set of scheduled appliance usages that actually got used plus the ones that were not scheduled but would leave the timeframe in the next step, and $(a_{a,t})_{i,1}$ is the power consumption of the i-th usage.

4. 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.

4.1. Thresholding Model

A quick subsection about the non machine learning baseline

4.2. Theoretical Optimum

A quick subsection about the theoretical optimum

4.3. Reinforcement Learning

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

5. Results

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

References

Nakabi, T. A. and Toivanen, P. Deep reinforcement learning for energy management in a microgrid with flexible demand. *Sustainable Energy, Grids and Networks*, 2020. URL https://doi.org/10.1016/j.segan.2020.100413.