Notes Week 51

Final reward function:

$$R(s, a, s') = 0.3 \cdot Comfort + 0.1 \cdot Emissions + 0.3 \cdot Grid + 0.3 \cdot Resilience \tag{1}$$

where:

$$Comfort = U, \tag{2}$$

$$Emissions = G, \tag{3}$$

$$Grid = \overline{R, L, D, A}, \tag{4}$$

$$Resilience = \overline{M, S}, \tag{5}$$

where these 4 reward components are made up of 8 key performance indicators (KPIs): carbon emissions (G), discomfort (U), ramping (R), 1 - load factor (L), daily peak (D), all-time peak (A), 1 - thermal resilience (M), and normalized unserved energy (S).

The grid and resilience reward components are averages over their KPIs.

All KPIs are given in notes Week 48, except that we have revised daily peak and all-time peak, which are calculated as follows:

Daily peak reward and All-time peak reward shaped as escalated rewards:

D: Daily peak

Maximum electricity consumption at any time step per day. Here, we have chosen to model this reward not as a delayed sparse reward at the end of the day, but as an escalated reward. The idea is to give reward proportional to the knowledge gained until that timestep while the accumulated reward will be the same as the reward otherwise received at the end of the day. As for the daily peak, at t = 0, we can say that the peak of that day will at least be the as big as the current consumption. Then, we compare this consumption to the following timestep, if there is a larger consumption, the difference is given as reward. If the energy demand from the grid at that timestep is lower, 0 reward is given.

$$D = D_{control} \div D_{baseline}, \tag{6}$$

where:

$$D_t = \begin{cases} max(0, E_t), & \text{if } (d \cdot h) = t \\ max(0, E_{d \cdot h}, E_{d \cdot h+1}, ..., E_t) - D_{t-1}, & \text{otherwise} \end{cases} \tag{7}$$

where:
E = Neighborhood-level net electricity consumption,
n = Total number of time steps, where every hour is one timestep,
t = Current time step,
d = Day,
h = Hours per day.

In this equation, we are considering all data of the current day up untill the current timestep. For each step, if the energy consumption is largest of that day, a reward is received. The accumulated reward at the end of the day will be the complete maximum electricity consumption of that day, but given to the agents in increments. This reward is to be minimized like the rest of the reward function.

Note: this part is averaged over all days in the episode when used in the original score calculation. In this step-wise conversion, the reward given to the agents is proportionally a lot larger than in the score calculation as the reward is not averaged. Perhaps this should be analyzed and weighted accordingly to better reflect the scoring function.

A: all-time peak

Here, taking the same approach will result in sparse rewards. Rewards will only be given to the agents when a new max energy consumption is achieved over the entire episode. Perhaps because this is a minimization problem, the zero sparse reward still gives the agent enough information as it is the optimal reward. Also, whenever a reward is gained in the episode it hopefully tells the agents to watch their energy consumption.

$$A = A_{control} \div A_{baseline}, \tag{8}$$

where:

$$A_t = \begin{cases} max(0, E_t), & \text{if } t = 0 \\ max(0, E_0, E_1, ..., E_t) - A_{t-1}, & \text{otherwise} \end{cases} \tag{9}$$

Note: this reward can also reconsidered.