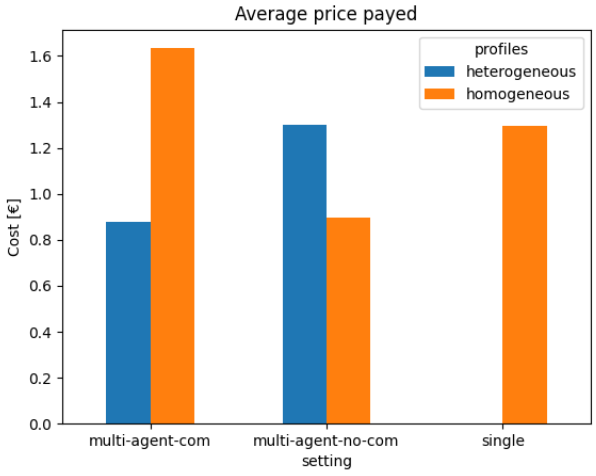
The following results have been obtained by using tabular Q-learning agent trained on 7 days of data. The results are from the validation day, the test day will be used at the very end.

I trained different models to see if using communication between agents was actually useful. For this I trained an agent on its own, to see if it learns the task appropriately and to have an idea of the cost it must pay for its consumption. A second scenario uses multiple agents that do not use each other's information in their Q-table. This means they can trade power among each other, but their decisions are not affected by any direct information exchange. The last scenario is the one where agents can trade power and this information is taken into account to take their decision. So the P2P exchange is part of the state space. In both multiagent settings there is a scenario where the agents have exactly the same load and PV generation, and one where these time series are different for each agent (the PV only differs in scale, since the agents are considered to be located next to each other).
For computational reasons (and also time) I've only used 2 agents and only trained on 7 days of data. I hope the results can still be viewed as relevant. With respect to the amount of data, more data will allow it to generalize better and work better for different months, but the principle should still hold. With respect to the number of agents, the learning algorithm has no knowledge of the number of peers, so this wouldn't change. The only thing that could happen is that there are larger variations, which would make it harder for the agent to learn.

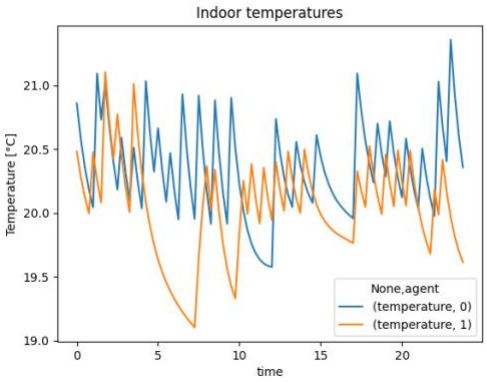
The first image shows the average costs of the agents for each scenario. In the heterogenous multi-agent setting the actors that can communicate do a better job. In the communicating homogenous case, one of the actors activates the heat pump too much at the end of the day, incurring large costs, which make the average shoot up. The other agent has a similar cost to the non-communicating agents.

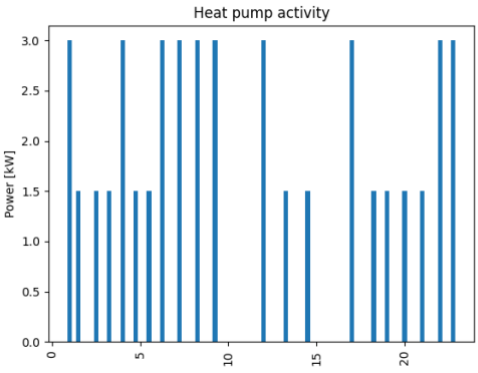


Considering that usually different agents will have different load and generation profiles, it seems beneficial to have explicit communication between them.

The following figures give a bit more insight into the agents' decisions for the communicating heterogeneous scenario. They show the agent manages to learn the correct task, deviating only within reasonable bounds. The temperature should not drop below 20°C, which does happen, but the agent does manage to get it back within bounds. The violations happen

during periods when the grid price is higher (not shown here, I'll make sure it's present in a following iteration). This could be improved by playing with the penalty for moving outside of the temperature comfort zone.





The last figure shows the loads and PV generation for the same situation. The values 0 and 1 in the legend are the agents' IDs