Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Best-Response Planning of Thermostatically Controlled Loads under Power Constraints

Frits de Nijs and **Matthijs T. J. Spaan** and **Mathijs M. de Weerdt**
{f.denijs, m.t.j.spaan, m.m.deweerdt}@tudelft.nl
Delft University of Technology, The Netherlands

Abstract

Renewable power sources such as wind and solar are inflexible in their energy production, which requires demand to rapidly follow supply in order to maintain energy balance. Promising controllable demands are air-conditioners and heat pumps which use electric energy to maintain a temperature at a setpoint. Such Thermostatically Controlled Loads (TCLs) have been shown to be able to follow a power curve using reactive control. In this paper we investigate the use of planning under uncertainty to pro-actively control an aggregation of TCLs to overcome temporary grid imbalance. We present a formal definition of the planning problem under consideration, which we model using the Multi-Agent Markov Decision Process (MMDP) framework.

Since we are dealing with hundreds of agents, solving the resulting MMDPs directly is intractable. Instead, we propose to decompose the problem by decoupling the interactions through arbitrage. Decomposition of the problem means relaxing the joint power consumption constraint, which means that joining the plans together can cause overconsumption. Arbitrage acts as a conflict resolution mechanism during policy execution, using the future expected value of policies to determine which TCLs should receive the available energy. We experimentally compare several methods to plan with arbitrage, and conclude that a best response-like mechanism is a scalable approach that returns near-optimal solutions.

Introduction

Reliable incorporation of decentralized renewable energy sources in power grids provides challenges for which Artificial Intelligence techniques can be employed. In order to take full advantage of wind and solar power, demand should follow their supply to maintain energy balance. The problem of how to influence the demand is called Demand Response, which we address using planning-under-uncertainty methodology. Planning to shift demands is most useful when storage capacity is available, offering flexibility by temporarily buffering energy. Unfortunately, large-scale electrical storage is unfeasible at the moment.

However, a large potential storage capacity can be found in the various heat buffers operated by consumers: houses, refrigerators and hot water reservoirs all need to be maintained at a certain temperature offset from the environment temperature, to which they decay over time. This requires constant action to counteract. The exact moment when these buffers are heated or cooled, however, can often be shifted. By storing more energy in the heat buffer now, we can later 'extract' this heat from the buffer in the form of reduced loads. In this sense, we may think of heat buffers controlled by thermostats as a kind of batteries (Hao et al. 2013). The storage potential in these batteries can be exploited through Demand Response.

These Thermostatically Controlled Loads (TCLs) can themselves be controlled automatically by the system operator. Existing work has mainly focused on using Control Theory to make an aggregation of TCLs closely follow an available power signal (Callaway 2009; Hao et al. 2013). There, control actions are applied every second, which requires the use of reactive systems. Their goal is to match both positive and negative fluctuations. This implies that the power available is always sufficient to cover demand, otherwise devices would begin to over- or under-perform.

Instead of using TCLs for balancing, we investigate the potential of TCLs for buffering temporary drops in power production, such as those caused by the fluctuations in renewable generation. These fluctuations typically occur on longer timescales (Koch, Zima, and Andersson 2009), potentially violating the consumers' comfort constraints if no preventive action is taken. To respond to fluctuations before they occur, we propose a planning approach with an objective to minimize the discomfort experienced by end-users.

In this paper we present a Power Constrained Planning (PCP) problem formulation, in which the deviation from the TCL setpoint is optimized under power availability constraints. For example for grids running in island mode, or when only (own) renewable energy may be used, the available power should be modeled as a constraint. We model this as a centralized planning problem with cooperative agents because in such scenarios the system operator and the TCL owners are both interested in a stable power supply, and hence their objectives align. Because in such cases power constraints may be too severe to guarantee comfort, we aim to *minimize* the total discomfort.

In order to solve the PCP problem, we model it as a Multi-agent Markov Decision Process (MMDP) (Boutilier 1996), in which each TCL is represented by an agent. However, as we are considering neighborhoods with hundreds of households, naively solving the resulting exponentially-sized MMDP will not suffice. In this paper, we show how we can exploit the fact that the only interaction among the agents is the allocation of available power. We decouple the problem and let each agent compute its own individual plan using a single-agent MDP. Combining the individual plans will result in a joint plan that might not be feasible due to power constraints. Hence, we introduce an arbitrage mechanism that ensures power constraints are respected by limiting the number of TCLs that can be switched on simultaneously. However, by in turn computing best-response policies and by learning the probability of being allowed to switch on, we compute scalable and effective solutions to the PCP problem.

We perform experiments on a small toy problem and on a larger test scenario. The toy problem allows us to clearly illustrate our ideas and to showcase the limited scalability of optimal centralized solutions. Using the larger scenario, we show that the best-response policies are able to scale to realistic problem sizes, while maintaining a solution quality that is near optimal.

Background

A thermostatic load is any device that is able to consume (electric) power for the heating or cooling of a body in relation to the outdoor temperature, such as refrigerators or central heating systems. The goal of the thermostat is to operate the device such that the temperature of the body remains as close as possible to a given setpoint at all times.

A Markov chain model of thermostats controlled by hysteresis controllers was presented by Mortensen and Haggerty (1988). In their model, the temperature of the body in the next time step $\theta_{i,t+1}$ is deduced from the current temperature $\theta_{i,t}$, the current outside temperature θ_t^{out}, a temperature input from the device θ_i^{pwr}, and a random temperature shift $\theta_{i,t}^{\text{rnd}}$ modeling exogenous actions such as opening a door. The hysteresis controller operates to keep the temperature within a deadband surrounding temperature setpoint $\theta_{i,t}^{\text{set}}$.

We use a similar model in this paper to make on/off decisions $m_{i,t}$ for every TCL i and for every discrete time period t. The length of such a period Δ, together with the thermal constants R_i (thermal resistance, °C / kW) and C_i (thermal capacitance, kWh / °C) determine how quickly the current temperature responds to the external factors through the fraction $a_i = \exp \frac{-\Delta}{R_i C_i}$, resulting in the model:

$$\theta_{i,t+1} = a_i \theta_{i,t} + (1 - a_i) \left(\theta_t^{\text{out}} + m_{i,t} \theta_i^{\text{pwr}}\right) + \theta_{i,t}^{\text{rnd}}. \quad (1)$$

Power Constrained Planning

We adapt the unconstrained model from the previous section into a Power Constrained Planning (PCP) problem. We use boldface characters to represent vectors of device parameters over all devices, i.e., $\boldsymbol{\theta}_t = [\theta_{1,t} \quad \theta_{2,t} \quad \cdots \quad \theta_{n,t}]$. Then, using the Hadamard product $\boldsymbol{b} = \boldsymbol{a} \circ \boldsymbol{\theta}_t \implies b_i = a_i \times \theta_{i,t} \quad \forall i$, we can define a state transition function to compute $\boldsymbol{\theta}_{t+1}$ as

$$f(\boldsymbol{\theta}_t, \boldsymbol{m}_t, \theta_t^{\text{out}}) = \boldsymbol{a} \circ \boldsymbol{\theta}_t + (1 - \boldsymbol{a}) \circ \left(\theta_t^{\text{out}} + \boldsymbol{m}_t \circ \boldsymbol{\theta}^{\text{pwr}}\right). \quad (2)$$

For simplicity, we keep random factor $\theta_{i,t}^{\text{rnd}}$ set to 0 for the planning phase, because we focus on distributions with expected value 0, and so that we can produce simpler optimal formulations.

With this function, we can define a planning problem using a given horizon h, the thermal properties of the n thermostatic loads with initial temperatures $\boldsymbol{\theta}_0$, the predicted outdoor temperature θ_t^{out}, and the predicted power constraint L_t. A solution is a device activation schedule that never switches on more devices than is allowed while minimizing cost function $c(\boldsymbol{\theta}_t)$. The entire planning problem becomes:

$$\begin{aligned} \underset{[\,\boldsymbol{m}_0 \;\boldsymbol{m}_1 \;\cdots\; \boldsymbol{m}_{h-1}\,]}{\text{minimize}} \quad & \sum_{t=0}^{h-1} c(\boldsymbol{\theta}_t) \\ \text{subject to} \quad & \boldsymbol{\theta}_{t+1} = f(\boldsymbol{\theta}_t, \boldsymbol{m}_t, \theta_t^{\text{out}}) \\ & \sum_{i=1}^{n} m_{i,t} \leq L_t \\ & m_{i,t} \in [0, 1] \qquad\qquad \forall i, t \end{aligned} \quad (3)$$

In this model power consumption is assumed to be constant within a time step, while in reality start-up transients occur when devices switch on. However, these occur only for a couple of seconds, after which consumption is constant, as is evident from measurements on heat-pump power consumption (van Lumig 2012, pages 25-28). We consider minute time scales, on which transients do not have a significant effect.

Due to the generality of the model, the controlled loads can have different objectives which can be expressed through the cost function. Besides typical functions such as the squared error $c(\boldsymbol{\theta}_t) = \sum_{i=1}^{n} \left(\theta_{i,t} - \theta_{i,t}^{\text{set}}\right)^2$ or maximum setpoint deviation $c(\boldsymbol{\theta}_t) = \max_i \left(\theta_{i,t} - \theta_{i,t}^{\text{set}}\right)$, we might imagine more application-specific functions. For example, a refrigerator may only incur high penalties when the temperature gets above a (thawing) threshold.

As a representative example of a cost function we use a variant of the squared error where minor (0.5 degrees) offsets incur no costs at all:

$$c(\boldsymbol{\theta}_t) = \sum_{i=1}^{n} \max\{0, \left|\theta_{i,t} - \theta_{i,t}^{\text{set}}\right| - 0.5\}^2. \quad (4)$$

We use this cost function because it has been shown in user studies that 90% generally acceptable comfort levels can be found in a small (2.5°C) band surrounding the ideal temperature (de Dear and Brager 1998). Additionally, it is implementable in mixed-integer programming constraints, and it imposes an intuitive notion of fairness; if even a single house is far from its setpoint, a large penalty is still incurred.

Optimal Solutions

To solve the PCP problem optimally, we can encode the formulation from Equation 3 as a mixed-integer program (MIP) with a minor reformulation. We introduce new variables $c_{i,t}$ and let $c(\boldsymbol{\theta}_t) = \sum_{i=1}^{n} c_{i,t}^2$. To ensure that these variables represent the correct cost values, we add the following constraints:

$$\begin{aligned} c_{i,t} &\geq 0 && \forall i,t \\ c_{i,t} &\geq \theta_{i,t} - \theta_{i,t}^{\text{set}} - 0.5 && \forall i,t \\ c_{i,t} &\geq -\theta_{i,t} + \theta_{i,t}^{\text{set}} - 0.5 && \forall i,t \end{aligned} \tag{5}$$

Alternatively, the problem can be modeled as a Multi-agent Markov decision process (MMDP) (Boutilier 1996). In our MMDP model we have a set of n agents that all have the same actions $\mathcal{A}$ available to them, and an agent-specific state $\mathcal{S}$. The transition function $T : \mathcal{S}^n \times \mathcal{A}^n \times \mathcal{S}^n \to [0, 1]$ describes for each joint state and joint action pair $(\mathbf{s}, \mathbf{t})$ the probability of attaining joint state $\mathbf{s}'$. The agents are cooperating to maximize the collective system reward. Agents are rewarded for their actions in a certain joint state, through the reward function $R : \mathcal{S}^n \times \mathcal{A}^n \to \mathbb{R}$.

Concretely, we propose the following MMDP model for the optimization problem defined in Equation 3.

The continuous temperature is discretized into k non-overlapping states s_j each defining a temperature interval $[\theta_{s_j,\text{min}}, \theta_{s_j,\text{max}})$. In addition to this, there are two extrema states s_{min} and s_{max} ranging from $(-\infty, \theta_{s_1,\text{min}})$ and $[\theta_{s_k,\text{max}}, \infty)$ respectively, resulting in the following state space of an agent: $\mathcal{S} = \{s_{\text{min}}, s_1, s_2, \ldots, s_k, s_{\text{max}}\}$.

Each agent has two actions available to it: $\mathcal{A} = \{\text{off}, \text{on}\}$. The previous action is not part of the state because short-cycling of the devices is already prevented by the sufficient time between decisions.

The rewards assigned to the agents in each time step are the costs depending on how large the deviation from the set-point $\theta_{i,t}^{\text{set}}$ is:

$$\sum_{i=1}^{n} -\max\{0, \left|s_i - \theta_{i,t}^{\text{set}}\right| - 0.5\}^2. \tag{6}$$

The transition function from a state s to a state s' is derived by applying the Markov heat-transfer function $f(\theta, m) = \mathbf{a}\theta + (1 - \mathbf{a})(\theta_{\text{outside}} + m\theta_{\text{heating}})$ to the lower and upper values of the temperature range $[\theta_{s,\text{min}}, \theta_{s,\text{max}})$. This produces a new range $[\theta'_{\text{min}}, \theta'_{\text{max}})$ that may overlap the ranges of multiple discrete states $s_j, s_{j+1}, \ldots$. The degree of overlap determines the (uniform) probability of transitioning to each of these potential future states. The factored (per agent) transition function is thus probabilistic, in order to be consistent with the heat-transfer function:

$$\begin{aligned} \theta'_{\text{min}} &= f(\theta_{s,\text{min}}, m) \\ \theta'_{\text{max}} &= f(\theta_{s,\text{max}}, m) \\ T(s, m, s') &= \frac{\min\left(\theta_{s',\text{max}}, \theta'_{\text{max}}\right) - \max\left(\theta_{s',\text{min}}, \theta'_{\text{min}}\right)}{\theta'_{\text{max}} - \theta'_{\text{min}}} \end{aligned} \tag{7}$$

The imposed power constraint L_t which limits the number of activated devices is then encoded in the joint transition function. The joint transition function specifies the cross product of all agents' action spaces $\mathcal{A}^n$. By removing those actions where the number of agents 'on' is more than L_t we obtain the required constraint.

The resulting MMDP is subsequently solved optimally using value iteration (Puterman 1994).

Decoupling, Arbitrage, and Best-Response

The only interaction among the TCLs is the allocation of the power availability. We therefore introduce a method that exploits this limited level of interaction as follows. First, we decouple the agents and let them find an individually optimal plan. Then we use an arbitrage mechanism upon execution to make sure that the agents together do not overuse the available power. Finally, we show how to let the agents coordinate their plans by simulating the arbitrage mechanism in advance, and performing a best-response to each other's policies. Each of these steps is discussed in more detail below.

First we completely *decouple* the agents, i.e., we factorize the MMDP by discarding the restriction L_t in the joint transition function. Then we solve a single-agent MDP for each agent separately.

Second, to overcome the problem that the agents' plans are no longer guaranteed to jointly stay within the power limit, we resolve the conflicts at policy execution using an *arbitrage mechanism*: in case too many devices want to switch on in a certain time step t, we iteratively search for the agent that expects to lose the least utility from switching off, and switch it off. This is repeated until the conflict is resolved. To determine which agent expects to lose the least utility by going from on to off we look at the difference between the planned utility scores in the value table.

Because this procedure is greedy and deterministic it risks getting caught in a local minimum. Thus, instead of always selecting the agent that expects to lose the least utility, we use the utility loss as a probability of being selected. A common probability distribution for fairly selecting between weighted alternatives is the logit equilibrium (McKelvey and Palfrey 1998).

Third, the plans for the single-agent MDPs take into account the effect of arbitrage and thereby indirectly of the plans of other agents by estimating how likely an agent is to be assigned such a constrained action. The following subsection describes how we arrive at good estimates of this probability.

Best-Response Planning for Arbitrage

The first step towards estimating the probability of being assigned energy is to take the pessimistic assumption that everyone always wants to make use of all available energy. Given this assumption, the probability of being able to activate the TCL is $p_{\text{on}}(t) = \frac{L_t}{n}$. This allows agents to detect potentially congested regions in time, but without knowing the actual level of congestion.

To estimate the actual demand in congested regions, we simulate policies computed in an earlier iteration and tally how much energy is requested. If energy is requested $\text{rq}_{\text{on},t}$

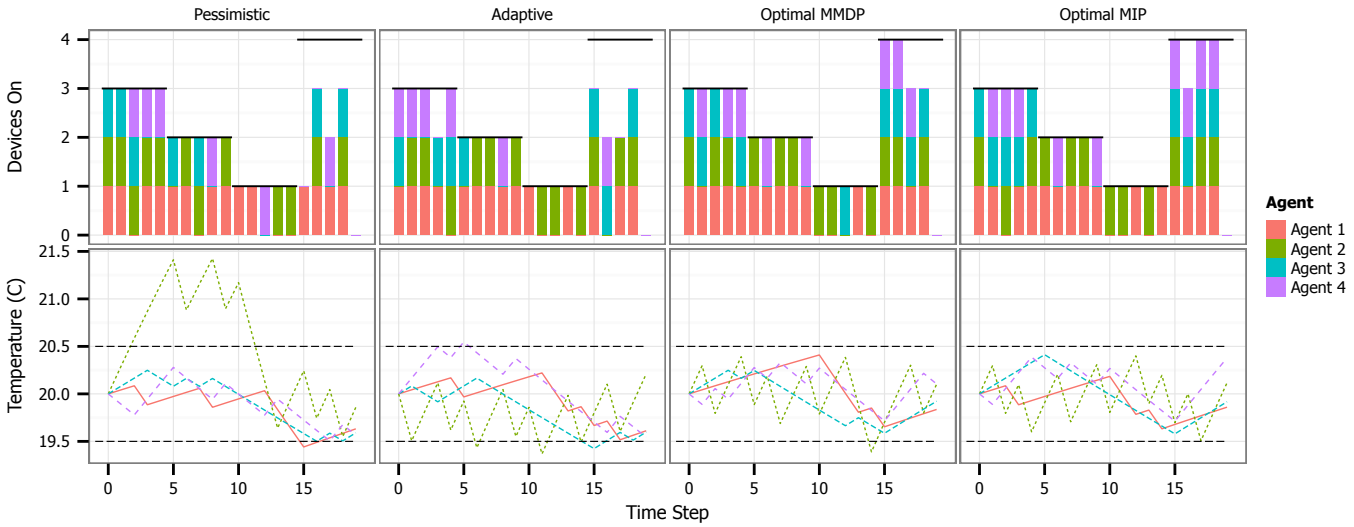

Figure 1: Activation scheme and agent temperatures in the example instance.

times in iteration i, then in the policy computation for iteration $i+1$ we assign probability $p_{\text{on}}(i,t) = \frac{L_t}{\text{rq}_{\text{on},i-1,t}}$, or $p_{\text{on}}(i,t) = 1$ if $\text{rq}_{\text{on},i-1,t} \leq L_t$.

Additionally, during periods where the expected demand is higher than the limit on the consumption, agents with a poor insulation level are much more likely to receive part of the available power. This is because those are the agents which incur a large potential error if they are skipped. Thus, agents that have a high degree of insulation are expected to make use of it by heating up before a constrained period. Such agents should learn that they will not receive power from arbitrage during constrained periods. To learn its relative probability of receiving power, each agent p_j counts $\text{rc}_{p_j,\text{on},i,t}$, the frequency with which his request for energy is granted in a time step t. Its probability then becomes $p_{\text{on}}(p_j,i,t) = \frac{\text{rc}_{p_j,\text{on},i-1,t}}{\text{rq}_{p_j,\text{on},i-1,t}}$.

Experimental Results on Small Instances

In order to demonstrate the value of the decomposition using arbitrage, we perform a number of experiments on small instances. These experiments are designed to show that:

1. Using individual probability to decompose the MMDP results in near-optimal performance on small instances.
2. Optimal solution methods do not scale beyond a handful of agents and a short horizon.

To investigate the behavior of the decomposition relative to the optimal solutions, we develop a toy instance with hand-crafted parameters resulting in a known optimal solution cost of 0. This instance has 4 agents, a horizon of 20 and $\theta^{\text{set}}_{i,t} = 20, \forall i,t$. In the first 5 time steps, 3 agents are allowed to switch on, in the next 5 time steps only 2 are allowed, followed by 5 time steps where only 1 is allowed. The final 5 time steps are unconstrained. Figure 1(top) illustrates the power constraints using black horizontal bars. Furthermore, the agents have insulation parameters chosen such that we know how frequently they need to switch on in order to keep the temperature within the deadband. In steady state agents 1 and 2 need to be on 75% of the time, while agents 3 and 4 need to be on 25% of the time.

We apply optimal MMDP and MIP solvers to this problem (with the MMDP creating plans for 9 distinct temperature states and $\theta_{\text{min}} = 18.5$, $\theta_{\text{max}} = 21.5$), as well as the arbitrage decompositions using the Pessimistic probability $p_{\text{on}}(t) = \frac{L_t}{n}$, and the decomposition using the Adaptive probability $p_{\text{on}}(p_j,i,t) = \frac{\text{rc}_{p_j,\text{on},i,t}}{\text{rq}_{p_j,\text{on},i,t}}$ (using 10 iterations, each with 10 simulations). The plans are subsequently evaluated using the planned decisions applied to the original continuous temperature function, Equation 1.

Because the optimal MMDP solver relies on a discretization of the temperature state, the level of discretization determines how likely it is to return optimal solution. Similarly, we expect the decompositions to not generate the optimal schedules. In particular, the pessimistic scheme is expected to strongly overheat several of the agents in the first 5 time steps, expecting heavy congestion in the subsequent middle 10 time steps. In contrast, the adaptive scheme should be able to detect that during the congested period agents 1 and 2 have much higher probability of switching on, promoting only agents 3 and 4 to pre-heat, limiting the error.

These expectations can be tested by looking at the cumulative error scores found in Figure 2 and the solutions generated by the four schemes in Figure 1. In Figure 1 we see that the optimal MIP is indeed able to keep all agents within the deadband simultaneously. It seems that the optimal MMDP suffers slightly from the low number of temperature states, which results in a small error w.r.t. optimal when agent 2 dips below 19.5 degrees in step 14.

Further, while the adaptive scheme stays close to the optimal, it is clear that the pessimistic approach is unable to find a good schedule. The reason for this can be seen in Figure 1: The second agent needs a lot of energy quickly, and

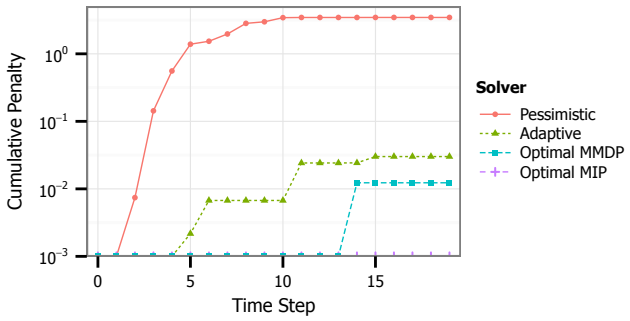

Figure 2: Cumulative penalty (log scale) in the example instance.

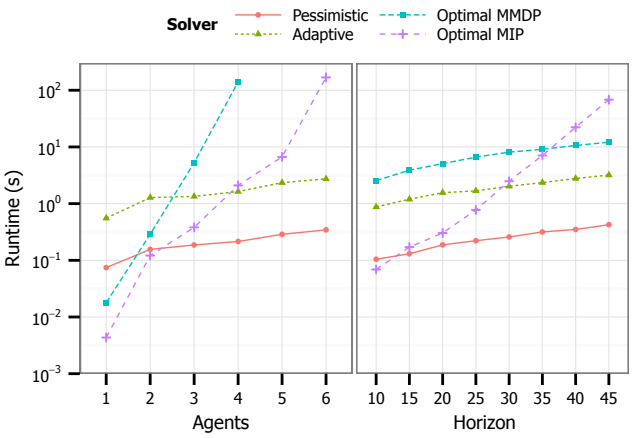

Figure 3: Scalability of the optimal and arbitrage methods for increasing agents and horizon, runtime on log scale.

expects it will be unlikely to receive it during the low power time steps. Therefore it strongly overheats before time 10. In contrast, the other methods are able to interleave agent 2 between the energy needs of the other agents. Hence, we conclude that we see all the expected behaviors.

For this small instance, all four considered formulations are able to return a solution within a reasonable amount of time (where solving the MMDP took the longest, at 12.5 minutes). However, we expect that the optimal formulations do not scale well, namely that their solving time grows exponentially with the instance size. On the other hand, the decoupled Pessimistic and Adaptive formulations are expected to grow in polynomial time, with the Adaptive formulation taking longer due to the iterative planning and simulation rounds.

We expect that both the MMDP and the MIP formulations grow exponentially in time with the number of agents, while we expect the MIP to additionally have exponential growth in the length of the horizon. The MMDP does not suffer from this because the Markovian property allows it to compute decisions iteratively, collapsing the computed decision sequences into a single expected value.

To evaluate the scalability we constructed a generator that builds instances which are similar to the example presented above, but with a variable number of agents and length of the horizon. We generated instances with the horizon fixed at 20 and the number of agents increasing from 1 to 6, and with the horizon from 5 to 45 and the number of agents fixed at 3 (10 instances per setting). In addition, we set the MMDP approach to use only 6 temperature states, and we imposed a run-time cut-off of 5 minutes. The average run-time performance is shown in Figure 3.

From the figure we can conclude that, indeed, the optimal methods are not scalable. Both methods appear to have run-time exponential in the number of agents, while the MIP also seems to have run-time exponential in the horizon.

Experimental Results on Real-World Instance

In the previous section, we show that the adaptive approach is able to stay close to the non-scalable optimal solutions in terms of solution quality. We expect that the adaptive approach is scalable to the size of typical real-world instances while maintaining near-optimal solution quality. To demonstrate this, in this section we apply it to a simulated neighborhood of 182 households equipped with heat-pumps. Their parameters are modeled to match measured performance (van Lumig 2012, page 60).

To determine if the solution quality is near optimal for this larger instance, we compare the pessimistic and the adaptive decomposition to a relaxation of the optimal MIP. This relaxation allows the devices to be switched on only partially. Because devices cannot be switched on partially, and because the decision time step granularity of 1 minute is the minimum to prevent short-cycling, it is not possible to implement the outcome of the relaxation. However, it is a lower bound on the optimal solution with binary activations.

In addition to this lower bound, we also compare our results with a hysteresis control solution that operates each house without look-ahead. For this control method we also perform arbitrage when required, by switching off units in order of their current temperature (hottest first).

For the instance, we model a two-day period during which a gradual decrease of the available power occurs starting from hour 6. At hour 20 the minimum capacity is reached and only 10 out of 182 heat pumps are allowed to be switched on. At the start of the second day, all devices can be switched on again. The households would all like to maintain an internal temperature of 21 degrees ($\theta^{\text{set}}_{i,t} = 21, \forall i, t$).

The decision frequency is set to once every minute. While it is unlikely that in a real-world scenario the power limit is known with such accuracy, this granularity allows each unit to switch just in time, and it also serves as a worst-case problem size to demonstrate scalability. Since the agents are now decoupled, we can solve the MMDP with much finer temperature discretization. For the adaptive decomposition we discretized temperature from 16 to 24 degrees over 80 states, resulting in bins of 0.1 degree width. Computed plans are again evaluated using the continuous temperature evolution (Equation 1) using a timescale of one second per time step (thus, a decision is held constant for 60 time steps in the simulation). Finally, in the simulation procedure the random component is enabled; for each house in every simulation

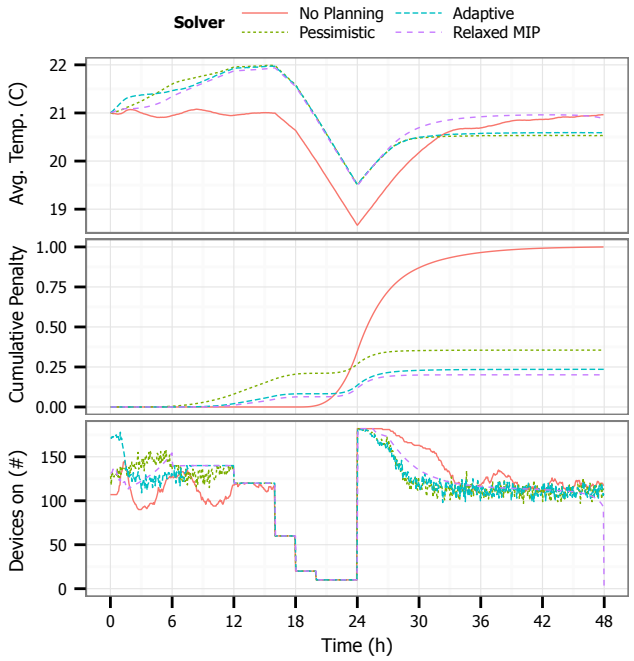

Figure 4: Average temperature, normalized cumulative penalty, and total load in the real-world instance.

step a value is drawn from the Gaussian distribution with mean 0 and standard deviation 0.01. Figure 4 presents the average indoor temperature, the normalized cumulative error and the number of devices switched on for this instance.

The cumulative penalty of the adaptive decomposition stays close to the MIP relaxation lower bound, which confirms our expectation that the adaptive decomposition stays close to optimal even in this larger instance. In addition we again see that the pessimistic decomposition heats up too much. This is evident from the cumulative penalty because the amount of error incurred above the deadband is higher than below it. Nevertheless, both planning solutions perform much better than the non-anticipatory control.

In addition, we can see in the graph of the average temperature (top) that the pessimistic decomposition starts heating earlier, which results in a larger number of devices switched on before the constraints kick in. In the graph showing devices switched on (bottom) we can also observe the gradual behavior of the relaxation versus the spiky discrete behavior of the decompositions.

Finally, it is important to note that computing the 182 policies for the adaptive decomposition took only 7.5 minutes, less time than it took the optimal MMDP solver to compute a solution for the four agent toy example. This demonstrates that the adaptive decomposition is indeed scalable to real-world instances.

Related Work

Mathieu et al. (2013) propose controlling thermostatic loads as arbitrage, essentially making them follow the inverse of a real-time price signal. They consider a control scenario where loads are only allowed to be controlled inside their deadbands, which is unobtrusive to the end-user. Their focus is on reducing the energy costs of consumers, under the assumption that sufficient power is available to maintain temperature inside the deadbands as required. They do not keep track of the states of individual devices in the planning and control phases, which causes a mismatch between their expected and actual performance, and limits applicability to relatively homogeneous collections of TCLs.

Stadler et al. (2009) present the dynamics of a system of setpoint controlled refrigeration units. Their results demonstrate that planning is necessary to get the most out of the storage capacity, which matches our observations.

A planning problem definition related to our own is presented by Bosman et al. (2012). The authors consider the planning of a fleet of MicroCHP units, so that their profit is maximized while the heat demand and minimum and maximum power production constraints are satisfied. The local search algorithm used first builds plans for each device individually using Dynamic Programming, which are then combined into a global plan and checked against the power constraints. If constraints are violated, the objective function of each house is updated and the devices replan until a feasible solution is found. Compared to our contribution the authors do not consider time-variable power constraints. Their heuristics are also much more computationally intensive, which limits their experiments to planning interval sizes and number of devices planned of 15 minute intervals over 24 hours and 100 units respectively, while requiring more than an hour of computation time.

Best-response planning has been used to coordinate routing and charging of electrical vehicles under uncertainty to minimize combined waiting and travel time (de Weerdt et al. 2013). Such behavior is quite natural for agents, and known to eventually converge to a stable situation in congested settings (Milchtaich 1996). This supports the idea that the presented approach may relatively easily transfer as well to situations with strategic agents.

Conclusions and Future Work

In this paper we investigate planning the activation of Thermostatically Controlled Loads (TCLs) for buffering for periods of low power availability. To this end we present a planning problem definition for the optimal control of Thermostatically Controlled Loads under power constraints. Because optimal solution methods do not scale on this problem definition, we propose a decoupling of the Multi-agent Markov Decision Process (MMDP) into an MDP per agent. At policy execution time, any conflicts that occur are resolved using an arbitrage mechanism, which greedily uses the policies to determine which agents benefit most from the available energy. By using a best-response planning process, agents are able to learn how likely they are to receive energy in a certain time step. From evaluating the adaptive decomposition on both large and small instances, we conclude that it is able to return near-optimal solutions while remaining scalable. This method therefore seems promising to control an aggregation of TCLs.

To extend this work we are considering two directions. On the one hand, we currently assume that the power production curve is known. In practice we do not know exactly how much power will be available for the TCLs in the future. Thus, we want to investigate planning TCLs under uncertainty. An advantage of the MDP framework is that it naturally allows for encoding uncertainty through the transition function, which makes the existing approach a promising starting point. On the other hand, we want to look into the addition of power quality constraints. Switching on many TCLs on a single feeder cable causes a large voltage drop for the last TCL in line. This can cause violations of the power quality norms, and risks damaging electric appliances. Hence, planning should be extended to take into account not just power but also voltage constraints in the grid.

Another interesting avenue of future work is the application of arbitrage and best-response to other planning problems. Because the MMDP framework considered is quite general, we expect the proposed techniques to apply as well to other problems where agents share a common infrastructure but otherwise do not have constraints in common.

Acknowledgments

Support of this research by network company Alliander is gratefully acknowledged.

References

Bosman, M.; Bakker, V.; Molderink, A.; Hurink, J.; and Smit, G. 2012. Planning the production of a fleet of domestic combined heat and power generators. *European Journal of Operational Research* 216(1):140–151.

Boutilier, C. 1996. Planning, Learning and Coordination in Multiagent Decision Processes. In *TARK*, 195–210.

Callaway, D. S. 2009. Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy. *Energy Conversion and Management* 50(5):1389–1400.

de Dear, R. J., and Brager, G. S. 1998. Developing an Adaptive Model of Thermal Comfort and Preference. *ASHRAE Transactions* 104(1):145–167.

de Weerdt, M. M.; Gerding, E.; Stein, S.; Robu, V.; and Jennings, N. R. 2013. Intention-aware routing to minimise delays at electric vehicle charging stations. In Rossi, F., ed., *Proceedings of the International Joint Conference on Artificial Intelligence*, 83–89. AAAI Press.

Hao, H.; Sanandaji, B. M.; Poolla, K.; and Vincent, T. L. 2013. A Generalized Battery Model of a Collection of Thermostatically Controlled Loads for Providing Ancillary Service. In *Annual Allerton Conference on Communication, Control and Computing*.

Koch, S.; Zima, M.; and Andersson, G. 2009. Active Coordination of Thermal Household Appliances for Load Management Purposes. In *Symposium on Power Plants and Power Systems Control*.

Mathieu, J. L.; Kamgarpour, M.; Lygeros, J.; and Callaway, D. S. 2013. Energy Arbitrage with Thermostatically Controlled Loads. In *European Control Conference*, 2519–2526.

McKelvey, R., and Palfrey, T. 1998. Quantal response equilibria for extensive form games. *Experimental economics* 1(1):9–41.

Milchtaich, I. 1996. Congestion games with player-specific payoff functions. *Games and economic behavior* 13(1):111–124.

Mortensen, R., and Haggerty, K. 1988. A stochastic computer model for heating and cooling loads. *IEEE Transactions on Power Systems* 3(3):1213–1219.

Puterman, M. L. 1994. *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. New York, NY: John Wiley & Sons, Inc.

Stadler, M.; Krause, W.; Sonnenschein, M.; and Vogel, U. 2009. Modelling and evaluation of control schemes for enhancing load shift of electricity demand for cooling devices. *Environmental Modelling & Software* 24(2):285–295.

van Lumig, M. 2012. Inventarisatie warmtepompmetingen WP1. Technical report, Laborelec.