2016 American Control Conference (ACC)
Boston Marriott Copley Place
July 6-8, 2016. Boston, MA, USA

Hierarchical Management of Demand Response Events with On/Off Loads

Andres Cortés and Sonia Martínez

***Abstract*— A suboptimal hierarchical approach for the management of demand response events (DRE) is proposed. We introduce a DRE model that explicitly accounts for on/off loads and leverages the inherent storage capacity of thermal loads provided by their inertia. Our approach can compute in a tractable time a feasible suboptimal solution to the DRE management problem, while user privacy is preserved. We present an MPC implementation of our approach and show the performance of our strategy in different DRE simulation scenarios.**

I. Introduction

The penetration of renewable generation into the power grid and the retirement of conventional generation from service poses significant operational challenges. Demand response, by which a virtual reserve capacity can be created using flexible loads, is a paradigm under investigation that can help overcome it.

We propose here a hierarchical demand response strategy, which is embedded in an MPC framework and which accounts for thermal and on/off loads. In this hierarchical approach, a single node (or some few nodes) acts as aggregators that collect information about loads and provides a coordination signal to them in order to achieve some system objectives. The larger computational effort is carried out in a decentralized way, while a central node provides a common coordination signal to all. Because the implementation is equivalent to that of a decentralized algorithm over a star-shaped topology, we note that this structure is widely qualified as 'decentralized' in the demand-response literature.

Demand response is currently the focus of a significant research effort. In [1], the authors present hierarchical algorithms for the dispatch of Distributed Energy Resources (DERs) and demand response. It computes the DER controls given a signal provided by an aggregator and by interacting with neighboring loads. In [2], the same authors address the DER control problem, by providing a fully distributed solution to an optimization problem to match the grid balance objective, without using an aggregator. Some works use the inherent storage capacity of some loads to plan demand response events minimizing the impact on users. In [3], the authors consider a hierarchical optimization approach for electric vehicle charging coordination, under usage constraints. In [4], [5], [6] the authors use a centralized model predictive control (MPC) formulation to take into account the thermal storage capacity of some loads and forecast information available at each time, to compute the load control. None of the aforementioned approaches consider on/off loads. In [7] the authors introduce a centralized MPC approach for thermal on/off loads, but they simply define a convex optimization problem and generate a on/off control using pulse width modulation (PWM). In our framework, doing so may lead to violating maximum power constraints. The paper [8] does consider on/off loads in the introduced setting, however, it is a centralized framework for a single household management. In [9], the load control problem is formulated for on/off loads, with the objective of minimizing the power generation cost. However, the authors do not explicitly address the integer constraints due to the on/off loads in their proposed solution. In [10], the authors present a framework that considers on/off loads, with an agile, but centralized suboptimal approach for load dispatch.

In this work, we introduce a hierarchical load control algorithm that explicitly takes into account the on/off nature of available loads, in order to provide a suboptimal control for a demand response event (DRE). Our formulation accounts for the energy storage capacity exhibited by thermal loads. To this end, we propose a thermal model that includes outside temperature as a disturbance. We formulate the problem as a mixed-integer program, with the objective of minimizing the effect of a demand response event (DRE) on the users' comfort. In order to solve the problem, we propose a greedy-like algorithm that provides a feasible solution with a reasonably good performance, using only computationally tractable methods. The introduced algorithm solves a convex relaxation of the original problem, and uses the relaxed control input for all on/off loads as a measure of need for power. After this, the resources are assigned using a greedy approach that provides power to the loads that need it the most until the maximum available power is allocated, via thresholding. These steps achieve a suboptimal solution of the mixed-integer problem. Both the convex optimization and the on/off load assignment via thresholding are carried out following the hierarchical architecture or, following the demand-response literature, a decentralized implementation over

This work has been supported by the NSF Award CMMI-1434819. A. Cortés and S. Martínez are with Department of Mechanical and Aerospace Engineering, University of California, San Diego,9500 Gilman Drive La Jolla, CA 92093-0411{andrescortes4500@gmail.com, soniamd@soe.ucsd.edu}.

a star-shaped topology.

We also present an MPC implementation of this algorithm. Finally, we use simulations to show the algorithm performance for different load scenarios.

II. Management of Demand Response Events

Demand-response events (DRE) result into the shaping of flexible power demand over a time horizon to provide different ancillary services to the power grid. We associate a DRE with a coordinated action of a large amount of power loads that modify their power consumption during a certain time lapse $\mathcal{T}$, in order to maintain the generation/demand balance in the power grid. The amount of power that the DRE must provide/withdraw from the grid is generally established from a transaction in an energy market by an aggregator [11].

In order to implement this, we introduce a DRE manager, which is an entity in charge of a large group of buildings with certain flexibility in their electric loads. The DRE manager aims to drive all loads into satisfying the DRE requirements, while minimizing its impact on the users' comfort. In this particular study, we consider thermal loads, such as air conditioners and heaters; memoryless loads, such as light bulbs; and non-flexible loads, that must be invariably active (or inactive) at certain times of the day. Moreover, most of these loads are on/off loads. The control strategy is designed to take advantage of the inherent energy storage capacity of thermal loads, and also of the prior knowledge of variables such as temperature or natural illumination, from a previously determined forecast process.

Another major interest in the computation of load control is the users' privacy. In general, users may not want to share their comfort model with the DRE manager, since it could include information on the user's habits (e.g., times when they are performing some activity). This is why a control strategy that can be computed in a hierarchical way, is a priority in the present work.

A. Modeling a Demand Response Event

Let us consider a DRE manager in charge of a set $I \triangleq \{1, \dots, N\}$ of buildings. Each building $i \in I$ contains five different types of loads. Let $\mathrm{Th}_{\mathrm{on/off}}(i)$ be the set of thermal on/off loads, $\mathrm{Th}_{\mathrm{curt}}(i)$ be the set of thermal curtailable loads, $\mathrm{L}_{\mathrm{on/off}}(i)$ be the set of memoryless loads, and finally, let $\mathrm{L}_{\mathrm{curt}}(i)$ be the set of memoryless curtailable loads for building $i \in I$. We denote by $L(i) \triangleq \mathrm{Th}_{\mathrm{on/off}}(i) \cup \mathrm{Th}_{\mathrm{curt}}(i) \cup \mathrm{L}_{\mathrm{on/off}}(i) \cup \mathrm{L}_{\mathrm{curt}}(i)$ as the set of flexible loads in $i \in I$.

A DRE time horizon $\mathcal{T}$ is divided into T time slots with duration $\Delta t = \mathcal{T}/T$; we let τ denote the sequence $\tau \triangleq \{0, \dots, T-1\}$ all discrete slots associated with it.

The power consumption of each flexible load $j \in L(i)$, $i \in I$, is denoted by $u_{ij}(t)$, where $t \in \tau$ is a discrete time instant. Since there is no feasible action for fixed loads in the buildings, we characterize them by a value $P_{\mathrm{fix}}(t)$.

A thermal load $j \in \mathbf{Th}(i) \triangleq \mathrm{Th}_{\mathrm{on/off}}(i) \cup \mathrm{Th}_{\mathrm{curt}}(i)$ is modeled by a discrete-time SISO linear system as follows:

$$\begin{aligned} x_{ij}(t+1) &= A_{ij}x_{ij}(t) + B^1_{ij}u_{ij}(t) + B^2_{ij}T^a_{ij}(t), \\ T_{ij}(t) &= C_{ij}x_{ij}(t), \\ x_{ij}(0) &= x^0_{ij}, \end{aligned} \tag{1}$$

where $x_{ij}(t)$ is the system state, $T_{ij}(t)$ is the temperature inside the room corresponding to the thermal load j, and $T^a_{ij}(t)$ is the outdoors temperature for the load at time $t \in \tau$. The vector $x_{ij}(0)$ represents the state at the beginning of the DRE. This discrete-time model may come from either an identification process using input-output data, or from the discretization of a continuous-time thermal model (e.g., an RC thermal model [12]) with time step Δt. Each load in a building has a discomfort value that is associated to its power input. For instance, the comfort value of a thermal load $j \in \mathbf{Th}(i)$ is given by:

$$\begin{aligned} f_{ij}(t) \triangleq & \kappa_{ij}(\max\{0, T_{ij}(t+1) - T^{\max}_{ij}\} \\ & + \max\{0, T^{\min}_{ij} - T_{ij}(t+1)\}), \end{aligned}$$

where $[T^{\min}_{ij}, T^{\max}_{ij}]$ is the temperature interval in which the users of the j^{th} thermal load in the i^{th} building are most comfortable, and κ_{ij} the users' tolerance to discomfort. Note the time shift in the temperature value, which is consistent with the fact that $T_{ij}(t+1)$ directly depends on $u_{ij}(t)$ for all $j \in \mathbf{Th}(i)$, $i \in I$. For memoryless loads we introduce a (generally nonnegative) discomfort function defined as $f_{ij}(t) = \alpha_{ij}(t)u_{ij}(t) + \beta_{ij}(t)$, where $\alpha_{ij}(t), \beta_{ij}(t) \in \mathbb{R}$, for all $t \in \tau$. The DRE itself is modeled as an upper (lower) bound on the amount of energy the whole set of buildings can use. This information is based on the load forecast on the power grid. For simplicity, we represent this bound by the constraints $\sum_{i=1}^{N}\sum_{j\in L(i)} u_{ij}(t) \le P_{\max}(t)$, for all $t \in \tau$.

Remark 2.1: In this study, we only consider DREs where a positive power compensation is required, i.e., the demand must be shortened. In the opposite case, the entire procedure can be adapted analogously. ◇

B. Optimal DRE control problem formulation

Here, we formulate an optimal control problem that results into the minimization of the general discomfort among the users of all the loads during the DRE. An algorithm to solve this problem is proposed in Section III.

The DRE manager will aim to solve the following optimization problem:

$$\mathcal{P}1: \text{minimize}_u \sum_{t=0}^{T-1}\sum_{i=1}^{N}\sum_{j\in L(i)} f_{ij}(t) \tag{2a}$$

subject to:

$$\text{Equation (1)}, \quad \forall j \in \mathbf{Th}(i), \forall i \in I, \forall t \in \tau, \tag{2b}$$

$$u_{ij}(t) \in [0, u^{\max}_{ij}], \quad \forall j \in \mathbf{Curt}(i), \forall i \in I, \forall t \in \tau, \tag{2c}$$

$$u_{ij}(t) \in \{0, u^{\text{on}}_{ij}\}, \quad \forall j \in \mathbf{Onoff}(i), \forall i \in I, \forall t \in \tau, \tag{2d}$$

$$\sum_{i=1}^{N}\sum_{j\in L(i)} u_{ij}(t) \le P_{\max}(t), \quad \forall t \in \tau. \tag{2e}$$

Notice that the state of the thermal loads at time $t = 0$ corresponds to the system state immediately prior to the beginning of the DRE.

The previous problem presents a convex cost function, however, constraints described by (2d) are binary, hence, the problem becomes is a mixed integer program. Since mixed integer programs are NP-complete, there is no algorithm that can solve it in polynomial time, and the solution time grows exponentially as the amount of integer variables grows.

III. Solution Approach

Privacy is a major objective for our load management solution. Thus, a centralized approach in which the DRE manager knows the model of all loads and discomfort functions of users may not be acceptable. Moreover, for large amounts of buildings or loads, the problem to be solved in a centralized way could grow too large to be manageable. Hence, the solution approach we consider must be susceptible of being executed in a hierarchical way, with parallel computation, where users only have to share with the DRE manager estimates of their demand profiles, but no comfort parameters, or usage patterns.

Our solution approach consists of two steps: convex optimization and thresholding. The thresholding step is devoted to use the result from the convex relaxation of $\mathcal{P}1$, that is not feasible to the problem and generate a feasible solution to it, in a greedy way, trying not to deteriorate the service provided to the users.

We describe the overall execution of these steps in the following, and leave the specific details for Subsection III-A and III-B. Then, we propose a hierarchical implementation of these steps in Section IV.

We consider the problem $\mathcal{P}2(0)$; see Subsection III-A, which is a convex relaxation of the problem $\mathcal{P}1$ in (2), with the only difference that we replace the constraints (2d) by $u_{ij}(t) \in [0, u_{ij}^{\text{on}}]$, for all on/off loads. Let $v^{\star,0}$ be an optimal solution of this relaxed problem. If we compute $y_{ij}^{\star,0}(t) \triangleq v_{ij}^{\star,0}(t)/u_{ij}^{\text{on}} \in [0,1]$, the result can be interpreted as the *level of urgency* that load j in building i has at time $t \in \tau$. For the sake of clarity, consider the time $t = 0$. The value of $y_{ij}(0)$ for all on/off loads can be used to establish the relative priority of these loads, and thus determine what loads should be on, based on the limited available power resources.

A threshold variable $\theta(0) \in [0,1]$ is introduced to decide on the state of on/off loads. Thus, all on/off loads for which $y_{ij}^{\star,0}(0) \in (0,1)$, $y_{ij}^{\star,0}(0) < \theta(0)$, must turn off, while those for which $y_{ij}^{\star,0}(0) \geq \theta(0)$ must turn on, i.e.:

$$\hat{u}_{ij}(0) = \begin{cases} u_{ij}^{\text{on}} & \text{if } y_{ij}^{\star,0}(0) \geq \theta(0) \\ 0 & \text{otherwise,} \end{cases}$$

for all $j \in \mathbf{Onoff}(i)$, $i \in I$, where $\hat{u}_{ij}(0)$ is defined as the control input to load $j \in L(i)$, $i \in I$. The threshold variable can always be chosen in such a way that after turning on and off the corresponding loads, the constraint on the maximum allowed demand is satisfied (see Lemma 3.1). Once the value for the on/off controllers $\hat{u}_{ij}(0)$ for all loads have been chosen, we proceed to solve the problem $\mathcal{P}2(0)$ as described above, but including the constraints $u_{ij}(0) = \hat{u}_{ij}(0)$, for all $j \in \mathbf{Onoff}(i)$, $i \in I$.

This computation will perform two tasks: i) to refine the computed values of $\hat{u}_{ij}$ for all loads $j \in \mathbf{Curt}(i)$, improving the use of resources at time $t = 0$, and ii) to provide a computation of the level of urgency for all on/off loads at time $t = 1$. Then, a threshold can be computed for $t = 1$, leading to the control values for all on/off loads at such time.

In this way, we increasingly fix the values of all on/off loads for each time $t \in \{1, \ldots, T-1\}$, given the previously computed control values control values for such loads at all times $q \in \{0, \ldots, t-1\}$.

A. Step 1: Convex optimization

More precisely, the proposed relaxation is defined next.

$$\mathcal{P}2(t) : \text{minimize}_u \sum_{q=0}^{T-1} \sum_{i=1}^{N} \sum_{j \in L(i)} f_{ij}(q)$$

subject to:

Equation (1), $\quad \forall j \in \mathbf{Th}(i), \forall i \in I, \forall q \in \tau,$

$u_{ij}(q) \in [0, u_{ij}^{\max}], \forall j \in \mathbf{Curt}(i), \forall i \in I, \forall q \in \{0, \ldots, T-1\},$

$u_{ij}(q) \in [0, u_{ij}^{\text{on}}], \forall j \in \mathbf{Onoff}(i), \forall i \in I, \forall q \in \{t, \ldots, T-1\},$

$u_{ij}(q) = \hat{u}_{ij}(q), \forall j \in \mathbf{Onoff}(i), \forall i \in I, \forall q \in \{0, \ldots, t-1\},$

$$\sum_{i=1}^{N} \sum_{j=1}^{L(i)} u_{ij}(q) \leq P_{\max}(q), \quad \forall q \in \{t, \ldots, T-1\}, \tag{3a}$$

for all $t \in \{1, \ldots, T-1\}$. Note that the problem $\mathcal{P}2(t)$ simply consists in relaxing the integer constraints for all on/off loads for all times $q \in \{t, \ldots, T-1\}$, fixing the previous computed control values for all loads at times $q \in \{0, \ldots, t-1\}$. Define $v^{\star,t} \triangleq \{v_{ij}^{\star,t}(q)\}_{j \in L(i), i \in I, q \in \tau}$ as the solution of the problem $\mathcal{P}2(t)$, for all $t \in \{1, \ldots, T-1\}$. This will be used to perform the thresholding procedure for the computed control values at time $t \in \tau$.

The following result establishes the existence of a suitable threshold to compute a feasible solution for all times $t \in \tau$.

***Proposition** 3.1:* Let $v^{\star,t}$ be a solution to the convex relaxation $\mathcal{P}2(t)$, for all $t \in \{0, \ldots, T-1\}$. Then, there exists a $\theta(t) \in [0,1]$ such that if:

$$\hat{u}_{ij}(t) = \begin{cases} u_{ij}^{\text{on}} & \text{if } y_{ij}^{\star,t}(t) \geq \theta(t) \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

the constraint $\sum_{i=1}^{N} \sum_{j \in L(i)} \hat{u}_{ij}(t) \leq P_{\max}(t)$ holds. $\diamond$

We omit the proof of all results in this paper due to space limitations, see [13] for details on proofs.

B. Step 2: Threshold Computation

So far, we have explained how the thresholding procedure can be used to satisfy the constraint on the maximum available power. Now, we establish a way of choosing $\theta(t)$ so that the overall cost is minimized.

In general, we consider that the maximum available power is a scarce resource that must be split among all loads. Then, we propose a *greedy* strategy in which the "optimal" threshold is chosen in the following way.

***Definition** 3.1:* An optimal threshold $\theta^\star(t)$ is one such that if we choose $\theta^{\text{new}}(t) \triangleq \max\{y_{ij}^{\star,t}(t) > \theta^\star(t), \forall y_{ij}^{\star,t}(t) \mid j \in \mathbf{Onoff}(i), i \in I\}$, and the thresholding process is carried out using $\theta^{\text{new}}(t)$, then the solution does not satisfy $\sum_{i=1}^{N}\sum_{j\in L(i)} \hat{u}_{ij}(t) \leq P_{\max}(t)$. ◇

Algorithm 1 Approximation algorithm

for $t = 0$ to $T-1$ **do**

- Compute $v^{\star,t}$ as an optimizer of $\mathcal{P}2(t)$.
- Compute threshold $\theta(t)$ and $\hat{u}_{ij}(t)$, for all $j \in \mathbf{Onoff}(i)$, $i \in I$, according to Equation (4).

end for

Compute $\hat{u}_{ij}(t)$ for all $j \in \mathbf{Curt}(i)$, $i \in I$, $t \in \tau$, by solving $\mathcal{P}1$ with the constraint $u_{ij}(q) = \hat{u}_{ij}(q)$, $q \in \{0,\dots,T-1\}$, $j \in \mathbf{Onoff}(i)$, $i \in I$.

***Remark** 3.1:* The following drawback may affect the solution performance. A large amount of on/off loads—possibly all of them—can result in an identical value of $y_{ij}^{\star,t}(t) = \bar{y}(t)$, for some $t \in \tau$. This means that if all loads are on and the solution is infeasible, the optimal threshold $\theta^\star(t)$ as introduced above is such that $\theta^\star(t) < \bar{y}(t)$, leading to all loads to be off at time t. A simple way to break the symmetry is the introduction of a small perturbation such that the value $y_{ij}^{\star,t}(t) = v_{ij}^{\star,t}(t)/u_{ij}^{\text{on}} + \epsilon_{ij}(t)$, where $\epsilon_{ij}(t)$ is a random variable with uniform distribution and very low variance. With this disturbance, the probability that two loads have exactly the same value $y_{ij}^{\star,t}(t)$ is zero, and the thresholding approach can be carried out without significant modification. Nevertheless, it is very unlikely that having many loads, a scenario like this occurs in practice. ◇

IV. Hierarchical approach

Our solution approach has been structured in a way that all computations are amenable to decentralization. This means that the calculation of the control inputs for all loads can be made by the agents associated to each building $i \in I$ as we describe next.

Recall that our solution approach consists of two separate steps, namely, i) solution of a convex relaxation and ii) thresholding. Then, we use two algorithms, one for each step, that are executed in an iterative fashion, via an information exchange between the building agents and the DRE manager. Figure 1 shows the communication structure and the information exchange of this network. At each algorithm, each agent $i \in I$ provides the DRE manager a usage signal d_i^k, while the DRE manager returns a coordination signal c^k which depends on d_i^k. In order to respect users' privacy, d_i^k does not include comfort parameters or load models.

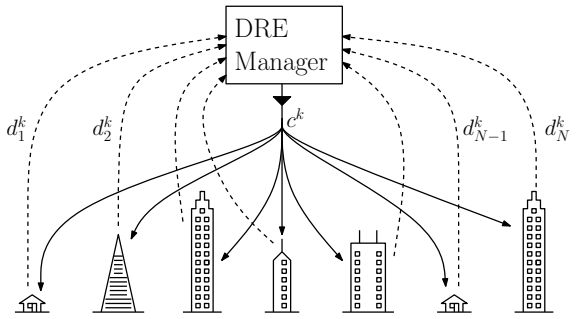

Fig. 1. Communication architecture for the hierarchical implementation of the algorithm. See Subsections IV-A IV-B for the definition of d_i^k and c^k in each case.

The definition of d_i^k and c_i^k will be introduced after the explanation of each step.

A. Step 1: Hierarchical convex optimization

In order to solve the optimization problem over the communication structure in Figure 1, we employ an augmented-Lagrangian methodology from [14] which is adapted to our setting. The Distributed Augmented Lagrangian Method (**ADAL**) of [14] is a provable-correct algorithm under the assumption of coupling equality constraints. In order to apply this algorithm with the same guarantees, we modify $\mathcal{P}2(t)$ as follows:

$$\mathcal{O}1: \text{minimize} \sum_{i=1}^{N+1} f_i(z_i)$$

subject to:

$$z_i \in Z_i, \forall i \in I \cup \{N+1\}, \tag{5a}$$

$$\sum_{i=1}^{N+1} F_i z_i = b, \tag{5b}$$

where each entry of the vector z_i corresponds to a decision variable associated to building i, either $\{x_{ij}(q), T_{ij}(q)\}$ for some load $j \in \mathbf{Th}(i)$, $q \in \tau$, $u_{ij}(q)$ for some load $j \in \mathbf{Curt}(i)$, $q \in \tau$, or $u_{ij}(q)$ for some load $j \in \mathbf{Onoff}(i)$, $q \in \{t,\dots,T-1\}$, for all $i \in I$. Likewise, $f_i(z_i) \triangleq \sum_{j\in L(i)}\sum_{q\in\tau} f_{ij}(q)$, with $f_{ij}(t)$ as defined in Subsection II-A, for all $i \in I$. After this, $Z_i \triangleq \{z_i \mid \text{Local constraints in } \mathcal{P}2(t) \text{ hold}\}$, for all $i \in I$, where Local constraints in $\mathcal{P}2(t)$ are given by all constraints of $\mathcal{P}2(t)$ except (3a). In addition, $F_i z_i \in \mathbb{R}^T_{\geq 0}$, corresponds to the vector with components $(F_i z_i)_\ell \triangleq \sum_{j\in L(i)} u_{ij}(q)$, where $(F_i z_i)_\ell$ is the ℓ^{th} component of $F_i z_i$. This implies that $F_i z_i$ is the aggregate demand profile of building i, given by the relaxed problem $\mathcal{P}2(t)$. By the non-negativity of $u_{ij}(q)$ for all j, q and $i \in I$, it is evident that $F_i z_i \succeq 0$, where the symbols $\succeq, \preceq$ indicate component-wise inequalities. Also, $b \in \mathbb{R}^T_{\geq 0}$ is such that the ℓ^{th} component of b corresponds to $P_{\max}(\ell-1)$. Notice that with these definitions, the inequality constraints in problem $\mathcal{P}2(t)$ correspond to $\sum_{i=1}^{N} F_i z_i \leq b$. The new variable $z_{N+1} \in \mathbb{R}^T_{\geq 0}$ is introduced simply as a slack variable to turn the inequality coupling constraints of $\mathcal{P}2(t)$

into equality constraints. Then, we define $f_{N+1}(z_{N+1}) = 0$, $F_{N+1}z_{N+1} = z_{N+1}$. Since $F_i z_i \succeq 0$ for all $i \in I$, it holds that $z_{N+1} \in Z_{N+1} \triangleq \{y \in \mathbb{R}^T \mid 0 \preceq y \preceq b\}$. Notice that the problem $\mathcal{P}1$ can be formulated as described above for $\mathcal{P}2(t)$. By the **ADAL** algorithm, agents and DRE manager execute the following iteration:

$$\hat{z}_i^k = \text{argmin}_{z_i \in Z_i} \mathcal{L}_i(z_i, \lambda^k) + \frac{\rho}{2}\|F_i z_i + \xi^k - F_i z_i^k\|^2,$$
$$z_i^{k+1} = (1-\gamma) z_i^k + \gamma \hat{z}_i^k, \tag{6}$$

for all $i \in I \cup \{N+1\}$, where $\mathcal{L}_i(z_i, \lambda) \triangleq f_i(z_i) + \lambda^\top F_i z_i$ and $\xi^k \triangleq \sum_{i=1}^{N+1} F_i z_i^k - b$, and:

$$\lambda^{k+1} = \lambda^k + \rho\gamma\xi^{k+1}, \tag{7}$$

where $\lambda^k \in \mathbb{R}^T$ is a Lagrange multiplier estimate.

Theorem 4.1: The algorithm in (6) - (7) converges to an optimal solution of $\mathcal{O}1$ for $0 < \gamma < 1/(N+1)$. ⋄

Now, let us describe the hierarchical implementation of the previous algorithm with the communication structure of Figure 1. At each iteration k, the values ξ^k, z_{N+1}^k, and λ^k are first computed by the DRE manager, which submits the signal $c^k = (\xi^k, \lambda^k)$ to buildings. After this, each building computes z_i^k and $d_i^k = F_i z_i^k$, for all $i \in I$. Then, d_i^k is sent to the DRE manager for the next iteration. Recall that $F_i z_i^k$ corresponds to the aggregate demand profile for the i^{th} building, for $i \in I$, which means that the users' privacy is preserved. The iteration is run until the constraint violation fulfills certain tolerance value.

B. Step 2: Hierarchical thresholding

In order to compute the threshold θ for time $t \in \tau$ in a hierarchical manner, we propose an iterative process as follows: first, the DRE manager sends an estimate of the optimal threshold for $\theta(t)$; then, based on the estimate and the solution of the relaxed optimization problem, the building agents compute the control inputs for all their on/off loads. Next, all buildings submit their aggregate load to the DRE manager, who updates the threshold estimate based on the latest information. The updating rule for the threshold estimate is given by:

$$x^{k+1}(t) = \begin{cases} \left[\frac{x_1^k(t)+x_2^k(t)}{2}, x_2^k(t)\right]^\top & \text{if } \sum_{i=1}^{N}\sum_{j\in L(i)} \hat{u}_{ij}^k(t) \le P_{\max}(t) \\ \left[x_1^k(t), \frac{x_1^k(t)+x_2^k(t)}{2}\right]^\top & \text{otherwise.} \end{cases}$$
$$\theta^k(t) = x_1^k(t), \tag{8}$$

with $x_1^0(t) = 1$ and $x_2^0(t) = 0$, for all $t \in \tau$. Recall that $\hat{u}_{ij}^k(t)$ is computed using the expression (4), with threshold $\theta^k(t)$, for all $j \in \mathbf{Onoff}(i)$. Observe this is a bisection-like search approaching asymptotically the optimal threshold $\theta^\star(t)$. From Proposition 3.1, we have that $\theta^0(t)$ provides a feasible solution for the optimization problem $\mathcal{P}1$. Furthermore, given the threshold recursion (8), it is easy to see that θ^k provides a feasible solution to the problem, for all $k \in \mathbb{N}$.

This algorithm can be run until the error $\|\theta^\star(t) - \theta^k(t)\| < \varepsilon$, for some $\varepsilon \ll 1$. Since this is a bisection-based algorithm, it is clear that $\|\theta^\star(t) - \theta^k(t)\| \le \|x_1^k(t) - x_2^k(t)\| \le (1/2)^k$, for all $k \in \mathbb{N}$. Thus, the stopping criterion can be recast as $k \ge -\log_2 \varepsilon$.

Following the communication architecture from Figure 1, in order to estimate $\theta^\star(t)$, for $t \in \tau$, we have that $c^k = \theta^k(t)$, while $d_i^k = \sum_{j\in L(i)} \hat{u}_{ij}^k(t)$.

Remark 4.1: Notice that the thresholding process aims to assign all the available power $P_{\max}(t)$ for the DRE at each time. This approach is not the best if the amount of power that all loads need at time t is less than $P_{\max}(t)$. Some thermal loads could be on in spite of being better off switched down. A simple way to solve this problem is to compute the power assignment using the introduced optimization/thresholding approach, and then using a lower-level local on/off controller for those thermal loads that were assigned power at time t. Such controller sets the input to zero if the load is hitting the upper bound in the comfort range of temperature. ⋄

V. Model Predictive Control Implementation

The model we use to construct the optimization problem $\mathcal{P}1$ is subject to several sources of uncertainty. The outside temperature $T_{ij}^a(t)$ comes from a forecast process that presents error. The thermal models themselves are not necessarily a perfect representation of the thermal loads. There can be unmodeled disturbances that affect the system performance. The parameters $\alpha_{ij}(t)$, $\beta_{ij}(t)$ may also come from forecast processes, e.g., if they are related to natural illumination in a room.

A way of addressing this uncertainty is via a Model Predictive Control (MPC) methodology [15]. By means of this, an optimization problem is solved at the beginning of each time slot $t \in \tau$, in which the initial conditions for the thermal systems are measured, and the forecast of those unknown variables is updated based on the latest information available at the moment. Then, from the computed control input for all time steps in $\{q \in \tau \mid q \ge t\}$, only the values corresponding to time t are applied on the plant. In order to compute the control input, we define the problem:

$\mathcal{M}1(t)$: minimize$_u$ $\sum_{q=t}^{T-1}\sum_{i=1}^{N}\sum_{j\in L(i)} f_{ij}(q)$
subject to:
Dynamics in Equation (1), with i. c. $x_{ij}(t) = x_{ij}^t$,
$\forall j \in \mathbf{Th}(i), \forall i \in I, \forall q \in \{t, \ldots, T-1\}$,
$u_{ij}(t) \in [0, u_{ij}^{\max}], \forall j \in \mathbf{Curt}(i), \forall i \in I, \forall q \in \{t, \ldots, T-1\}$,
$u_{ij}(t) \in \{0, u_{ij}^{\text{on}}\}, \forall j \in \mathbf{Onoff}(i), \forall i \in I, \forall q \in \{t, \ldots, T-1\}$,
$\sum_{i=1}^{N}\sum_{j\in L(i)} u_{ij}(t) \le P_{\max}(t), \quad \forall q \in \{t, \ldots, T-1\}$,

for each $t \in \tau$, where i.c. stands for initial conditions, and they are measured (or estimated) using the measured information at time $t \in \tau$. A feasible solution $\hat{u}^t$ of this problem can be computed in a hierarchical way using the methodology presented in Section IV. Moreover, since we only use the first step of the control input for each load, i.e., $\hat{u}_{ij}^t(t)$ into the system, we do not need to run all the computations described in Section IV. At each time $t \in \tau$, we simply need to:

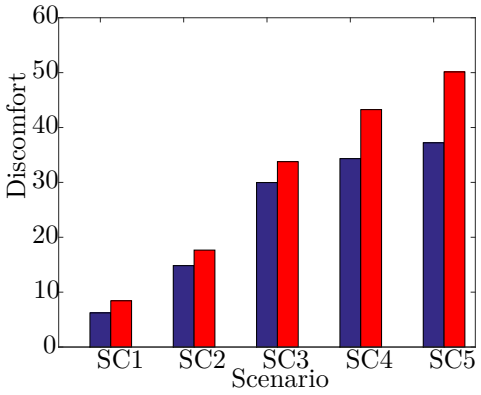

Fig. 2. In blue: discomfort (cost) for the solution to the convex relaxation of the problem $\mathcal{P}1$ (unfeasible). In red: normalized discomfort for solution to $\mathcal{P}1$ provided by our proposed method (feasible). Simulation over 5 different scenarios.

- Solve the convex relaxation of the problem $\mathcal{M}1$.
- Run a thresholding process to compute $\hat{u}_{ij}(t)$ for all $j \in \mathbf{Onoff}(i)$, $i \in I$.
- Refine the solution for all $\hat{u}_{ij}(t)$, $j \in \mathbf{Curt}(i)$, $i \in I$, by solving the convex relaxation of the problem $\mathcal{M}1$, including the constraint $u_{ij}(t) = \hat{u}_{ij}(t)$, for all $j \in \mathbf{Onoff}(i)$, $i \in I$.

VI. Simulations

In this section, we aim to show and discuss the performance of our hierarchical technique for demand response management. We run randomly generated scenarios with 10, 15, 20, 25 and 30 buildings, where each building corresponds to an average load of 0.2 MW. For each scenario, it is considered that the maximum available power for the entire set of buildings at each time is 30% of the maximum flexible load of the scenario. Figure 2 shows in red the cost (discomfort) for different scenarios for our suboptimal solution approach, while the blue bars represent a lower bound in the optimal cost. This lower bound consists in the solution to the convex relaxation of $\mathcal{P}1$ for each scenario. Recall that the solution to the convex relaxation of $\mathcal{P}1$ is not feasible for controlling on/off loads, while our approach does provide a feasible control for on/off loads. Hence the solution to the convex relaxation of $\mathcal{P}1$ is only used as a benchmark. It can be seen that the cost of the suboptimal solution drifts apart from to the convex lower bound as the amount of on/off loads increase.

VII. Conclusions

We propose a hierarchical method for coordination of loads in a demand response event (DRE). The method takes explicitly into account the fact that some loads are on/off, and aims for the suboptimal solution of a mixed-integer program. The objective of this problem is the minimization of user discomfort due to the DRE. The solution approach consists in solving a convex relaxation of the mixed-integer program, combined with a hierarchical greedy dispatch of the available power. While the convex relaxation provides a measure of the urgency level for power of each load in the DRE, a DRE manager must compute an urgency threshold to decide on the on/off loads that must receive power at each time. Additionally, we present an MPC implementation of our approach, in order to mitigate uncertainties and disturbances of the model. Simulations show that the loss of optimality of our approach is acceptable, and the possibility of computing the solution in a tractable time makes it a tool that can be used for problems with similar structure.

As future directions, we aim to establish analytic suboptimality bounds for our approach.

References

[1] A. Domínguez-García and C. N. Hadjicostis, "Distributed algorithms for control of demand response and distributed energy resources," in *IEEE Int. Conf. on Decision and Control*, (Florida, USA), pp. 27–32, December 2011.

[2] A. Domínguez-García, S. Cady, and C. Hadjicostis, "Decentralized optimal dispatch of distributed energy resources," in *IEEE Int. Conf. on Decision and Control*, pp. 3688–3693, 2012.

[3] A. Cortés and S. Martínez, "Optimal plug-in electric vehicle charging with schedule constraints," in *Allerton Conf. on Communications, Control and Computing*, pp. 262–266, 2013.

[4] B. Biegel, P. Andersen, T. Pedersen, K. Nielsen, J. Stoustrup, and L. Hansen, "Electricity market optimization of heat pump portfolio," in *IEEE Conf. on Control Applications*, pp. 294–301, 2013.

[5] R. Pedersen, J. Schwensen, B. Biegel, J. Stoustrup, and T. Green, "Aggregation and control of supermarket refrigeration systems in a smart grid," in *IFAC World Congress*, pp. 9942–9949, 2014.

[6] F. Tahersima, P. Andersen, and P. P. Madsen, "Economic energy distribution and consumption in a microgrid part1: Cell level controller," in *IEEE Conf. on Control Applications*, pp. 308–313, 2013.

[7] M. Avci, M. Erkoc, A. Rahmani, and S. Asfour, "Model predictive HVAC load control in buildings using real-time electricity pricing," *Energy and Buildings*, vol. 60, pp. 199–209, 2013.

[8] A. Agnetis, G. Dellino, P. Detti, G. Innocenti, G. de Pascale, and A. Vicino, "Appliance operation scheduling for electricity consumption optimization," in *IEEE Int. Conf. on Decision and Control*, pp. 5899–5904, 2011.

[9] L. Gkatzikis, T. Salonidis, N. Hegde, and L. Massoulié, "Electricity markets meet the home through demand response," in *IEEE Int. Conf. on Decision and Control*, pp. 5846–5851, 2012.

[10] B. Biegel, P. Andersen, T. Pedersen, K. Nielsen, J. Stoustrup, and L. Hansen, "Smart grid dispatch strategy for on/off demand-side devices," in *European Control Conference*, pp. 2541–2548, 2013.

[11] Q. Wang, C. Zhang, Y. Ding, G. Xydis, J. Wang, and J. Østergaard, "Review of real-time electricity markets for integrating distributed energy resources and demand response," *Applied Energy*, vol. 138, pp. 695–706, 2015.

[12] Y. Ma, G. Anderson, and F. Borrelli, "A distributed predictive control approach to building temperature regulation," in *American Control Conference*, pp. 2089–2094, IEEE, 2011.

[13] A. Cortés and S. Martínez, "Hierarchical management of demand response events with on/off loads." Preprint available at http://fausto.dynamic.ucsd.edu/andres/publications.html.

[14] N. Chatzipanagiotis, D. Dentcheva, and M. Zavlanos, "An augmented Lagrangian method for distributed optimization," *Mathematical Programming*, pp. 1–30, 2013.

[15] E. Camacho, T. Samad, M. Garcia-Sanz, and I. Hiskens, "Control for renewable energy and smart grids," *The Impact of Control Technology, Control Systems Society*, pp. 69–88, 2011.