Forecasting the Price-Response of a Pool of Buildings via Homothetic Inverse Optimization

Ricardo Fernández-Blanco, Juan Miguel Morales, Salvador Pineda

Abstract—This paper focuses on the day-ahead forecasting of the aggregate power of a pool of smart buildings equipped with thermostatically-controlled loads. We first propose the modeling of the aggregate behavior of its power trajectory by using a geometric approach. Specifically, we assume that the aggregate power is a homothet of a prototype building, whose physical and technical parameters are chosen to be the mean of those in the pool. This allows us to preserve the building thermal dynamics of the pool. We then apply inverse optimization to estimate the homothetic parameters with bilevel programming. The lower level characterizes the price-response of the ensemble by a set of homothetic parameters with bilevel programming. The lower level characterizes the price-response of the ensemble by a set of homothetic parameters with bilevel programming. The lower level minimizes the mean absolute error over a training sample. This bilevel program is transformed into a regularized nonlinear problem in which the homothetic parameters stemming from a heuristic are fixed, and the remaining variables initialized to the values resulting from that heuristic. This heuristic consists in solving two linear programs and its solution is deemed a suitable proxy for the original bilevel problem. The results have been compared to state-of-the-art methodologies.

Index Terms—Bilevel programming, data-driven approach, forecasting, homothet, inverse optimization, smart buildings.

NOMENCLATURE

The main notation used throughout the text is stated below for quick reference. Matrices are defined in bold and upper-case, vectors are indicated in bold and lower-case, superscript ' means observed, and symbol ~ refers to an estimated parameter. Other symbols are defined as required.

A. Sets and Indices

\( B \) Set of blocks, indexed by \( b = 1 \ldots n_B \).
\( D \) Set of days, indexed by \( d \).
\( H \) Set of hours, indexed by \( h = 1 \ldots n_H \).
\( \Omega \) Set of physical- and technical-related parameters.
\( \Phi \) Set of model-related parameters.
\( \mathbb{B} \) Feasible region of a building model.

B. Parameters

\( a_1 \) Energy dissipation rate.
\( a_2 \) Parameter defining the product of \( \eta \cdot R \).
\( c^a \) Discomfort cost.
\( C \) Thermal capacitance.

\( m_{b,h,d} \) Marginal utility of block \( b \) at hour \( h \) and day \( d \).
\( P \) Rated power of a thermostatically-controlled load.
\( R \) Thermal resistance.
\( z_{r,h,d} \) Regressor \( r \) at hour \( h \) and day \( d \).
\( \eta \) Coefficient of performance of a thermostatically-controlled load.
\( \theta_{0,d} \) Initial indoor air temperature in day \( d \).
\( \theta_r \) User-specified temperature set-point.
\( \theta_{amb,h,d} \) Outdoor air temperature at hour \( h \) and day \( d \).
\( \delta \) Discretization period.
\( \Delta \) Half of the temperature deadband.
\( \beta \) Scaling factor of the homothetic transformation.
\( \tau_h \) Translation factor of the homothetic transformation at hour \( h \).
\( \lambda_{h,d} \) Electricity price at hour \( h \).

C. Decision Variables

\( p_{h,d} \) Power of a thermostatically-controlled load at hour \( h \) and day \( d \).
\( s_{h,d} \) Slack variable for the temperature-related constraints at hour \( h \) and day \( d \).
\( \theta_{h,d} \) Indoor air temperature at hour \( h \) and day \( d \).

D. Matrices and Vectors

\( A, B \) Matrices associated with the matricial form of the building’s discrete dynamics.
\( c_d \) Vector representation of the building’s initial conditions in day \( d \).
\( t_d \) Vector representation of the component of the building’s discrete dynamics associated with the ambient temperature in day \( d \).
\( \Lambda \) The inverse of matrix \( A \).

I. INTRODUCTION

Distributed energy resources (DERs), such as distributed generators, electric vehicles, battery energy storage systems or demand response programs, are constantly growing every year and play a crucial role in the provision of multiple benefits to the power system [1]. In this paper, we focus on a recently popular DER, namely, an ensemble of smart buildings. This pool of buildings may efficiently utilize their thermal capacity while keeping the indoor air temperature at user-defined comfort levels in order to provide some degree of flexibility to the power system, by shifting their load in time or reducing the peak demand. In addition, this flexibility may allow its participation in a day-ahead electricity market or could even be viewed as a non-wire alternative to capacity expansion. However, as with any load in the electricity system,
its prediction is key to fully exploiting the benefits that can bring to the power system operation and planning [2].

Load forecasting has been extensively studied in the technical literature by using a plethora of methods such as autoregressive models with exogenous inputs [3] or neural networks [4]. However, all those models neglect the nature of the load to be predicted, e.g. thermostatically-controlled loads and electric vehicles are governed by different technical and physical constraints. Recently, the authors in [5] devised a novel inverse optimization (IO) approach to statistically forecast the aggregate load of a pool of price-responsive buildings in an hour-ahead setting. In that paper, they characterize the response of the load to the price by means of an optimization problem. The limitations of the model proposed in [5] are threefold: (i) the methodology is based on heuristics, (ii) the optimization models are tailored to single-step forecasts, and thus its use for multi-step forecasting is inappropriate, and, as a consequence, (iii) the building thermal dynamics are disregarded in the forecasting process.

As done in [5], we apply IO to forecast the aggregate response to the electricity price of a pool of buildings. However, our goal is to predict it in a day-ahead framework while also incorporating the building thermal dynamics into the optimization process. The goal of an IO problem is to infer the optimization model parameters given a set of observed decision variables or measurements collected by an observer. Recent advances on IO can be found in [6]–[9], and references therein. IO has also been applied to characterize price-responsive consumers in [10] and [11]. Saez-Gallego et al. [10] proposed an IO approach by using bilevel programming to infer the market bid parameters of a pool of price-responsive households such as step-wise marginal utility functions, maximum load pick-up and drop-off rates, and maximum and minimum power consumption bounds. Although [10] accounts for a refined model of the aggregate load of the households by including the ramping rates, it still neglects the thermal inertia governing the households’ decisions. Lu et al. [11] applied IO to estimate the demand response characteristics of price-responsive consumers, as similarly done in [10], and thus sharing the same shortcomings. Finally, reference [12] described a data-driven method to empirically estimate a robust feasible region of a pool of buildings. However, the thermal dynamics of the buildings were once again ignored from the estimation procedure.

One of the main contributions of this paper is the application of a geometric approach, i.e. we resort to the concept of homothety, to characterize the price-response of the ensemble of buildings for forecasting purposes. A homothety is a spatial transformation of an affine space. Hence, we assume that the feasible region of a pool of buildings can be represented as a homothet of a chosen prototype building by means of a dilation factor and a translation vector, namely the homothetic parameters. The homothetic representation of an aggregate of buildings has been first proposed in [13]. Specifically, Zhao et al. [13] put forward the modeling of the aggregate flexibility of a pool of thermostatically-controlled loads by using a geometric approach. The thrust of that paper was to derive sufficient and necessary virtual battery models that can be approximated by homothets of a virtual battery prototype. The authors demonstrated the benefits of such homothetic representation in the provision of flexibility for regulation services. To the best of our knowledge, this is the first time that a homothetic representation of an aggregate load has been applied for forecasting purposes. Consequently, we only rely on the estimation of the homothetic parameters to shape the aggregate feasible region of the pool, thus considerably reducing the computational complexity of the estimation algorithm and avoiding an undesirable overfitting. This work contributes to the technical literature as follows:

- From a modeling perspective, we propose a novel day-ahead forecasting technique for a pool of buildings via homothetic inverse optimization. The aggregate price-response is characterized by a set of marginal utility curves and a homothet of a prototype building. As novel distinctive features, this geometric approach endogenously accounts for the aggregate building thermal dynamics and allows us to solely rely on the estimation of two homothetic parameters and a set of marginal utility curves. We then apply IO to infer them through a given forecasting problem mimicking the price-response of the pool. Our approach, therefore, drastically reduces the complexity of the price-response model to be statistically estimated, while still capturing the thermal dynamics of the ensemble of buildings.
- The application of IO gives rise to a bilevel programming problem. We then propose the transformation of this bilevel program into a regularized nonlinear model which can be readily solved by nonlinear commercial solvers. To avoid meaningless local optimal solutions, we fix the homothetic parameters resulting from an efficient data-driven heuristic estimation procedure. From those heuristics, we also obtain initial values for the remaining variables. This heuristic estimation procedure is a suitable proxy for the solution of the original bilevel problem.
- The proposed forecasting technique has been compared with existing methodologies emphasizing its benefits for different degrees of heterogeneity among buildings.

The paper is outlined as follows. In Section II, we present a detailed derivation of the feasible region of the power trajectory of a smart building equipped with a thermostatically-controlled load. Section III provides the derivation of the feasible region for a pool of buildings by using a homothet of a building prototype. In Section IV, we provide (i) the proposed IO methodology based on a bilevel program and (ii) the heuristics we use to warm-start it and guide its solution. Section V describes the forecasting methodologies used to benchmark our proposal. Section VII provides insightful results. Finally, conclusions are duly drawn in Section VII.

II. INDIVIDUAL BUILDING MODEL

This section is devoted to deriving a mathematical model for a single smart building. Throughout the manuscript, we assume that variables and parameters without any superscript

1Also known as forward problem in the jargon of inverse optimization.
refer to building $i$, unless specified otherwise. Let us assume that we model the building $i$ as a single thermostatically-controlled load characterized by the heat transfer coefficient between the room air and the ambient, $U/A$, and the thermal capacitance of the room air, $C$. The electrical model of its thermal dynamics is represented in Fig. 1 wherein the resistance $R = 1/U_A$. In addition, we assume the building is equipped with a cooling system with a rated power $P$ and a coefficient of performance $\eta$. Applying the first Kirchhoff’s law, we obtain a first order differential equation at hour $h$:

$$C \frac{d\theta_h}{dh} + \frac{\theta_h - \theta_i^{amb}}{R} + \eta p_h = 0, \ \forall h \in \mathcal{H}. \quad (1)$$

The differential equation can be discretized with the Euler’s method. Thus the indoor air temperature can be expressed as:

$$\theta_h = a_1 \theta_{h-1} + (1-a_1) \left[ \theta_i^{amb} - a_2 p_h \right], \ \forall h \in \mathcal{H}, \quad (2)$$

where parameters $a_1 = 1 - \frac{\delta}{RC}$ and $a_2 = \eta R$, being $\delta$ the discretization period.

Bearing in mind both the temperature comfort bounds by the building’s occupants and the technical power limits of the cooling device, the feasible region of a single building for $n_H$ time periods within a day can be mathematically expressed as:

$$\theta_h = a_1 \theta_{h-1} + (1-a_1) \left[ \theta_i^{amb} - a_2 p_h \right], \ \forall h \in \mathcal{H} \quad (3a)$$

$$\theta_h \leq \theta_i \leq \theta_h, \ \forall h \in \mathcal{H} \quad (3b)$$

$$p_h \leq p_h \leq \bar{p}_h, \ \forall h \in \mathcal{H} \quad (3c)$$

where $\theta_i$ is the user-specified temperature set-point and $\Delta$ the half of the temperature deadband. In addition, $\bar{p}_h = 0$ and $\bar{p}_h = P$.

Summing up, the set of physical and technical parameters of building $i$ is $\mathcal{O}^i = \{R, C, \theta_i, \Delta, \eta, \theta_0, P\}$.

Conveniently and following the notation in [13], we can recast the thermal model (3) in matricial form:

$$\begin{align*}
\mathbf{p} \leq \mathbf{p} & \leq \bar{\mathbf{p}}, \\
\theta & \leq \Lambda \mathbf{B} \mathbf{p} + \Lambda \mathbf{c} + \Lambda \mathbf{t} \leq \bar{\theta},
\end{align*} \quad (4)$$

where $\Lambda$ is the inverse of $\mathbf{A}$: $\mathbf{A} = \mathbf{I}_{n_H} + \text{diag}(-a_1; -1)$, where $\mathbf{I}_{n_H}$ is the identity matrix of dimension $n_H$ and $\text{diag}(-a_1; -1)$ is a matrix of dimension $n_H$ with values $-a_1$ on the lower subdiagonal; $\mathbf{B} = -a_2 (1-a_1) \mathbf{I}_{n_H}$; $\mathbf{c}$ is the vector of initial conditions $[a_1 \theta_0, 0, ..., 0]^T$, being superscript $T$ the transpose operator; and $\mathbf{t}$ is the vector related to the ambient temperature, i.e., $\theta_i^{amb} (1-a_1)$. The $i$th component of $\mathbf{p}$, $\bar{\mathbf{p}}$, $\bar{\theta}$, $\bar{\theta}_h$, $\bar{p}_h$, respectively.

Finally, the feasible region of a thermal model for building $i$ can be written in a compact way as follows:

$$\mathbf{B} = \{ \mathbf{p} \in \mathbb{R}^{n_H} | \mathbf{F} \mathbf{p} \leq \mathbf{h} \}. \quad (5)$$

Assuming that the notation $(\mathbf{X}, \mathbf{Y})$ denotes the matrix $[\mathbf{X}^T, \mathbf{Y}^T]^T$ for two matrices $\mathbf{X}$ and $\mathbf{Y}$ with the same number of columns, the matrix $\mathbf{F} = (\mathbf{I}_{n_H}, -\mathbf{I}_{n_H}, \Lambda \mathbf{B}, -\Lambda \mathbf{B})$ and vector $\mathbf{h} = (\bar{\mathbf{p}}, -\bar{\mathbf{p}}, \bar{\theta} - \Lambda \mathbf{c} - \Lambda \mathbf{t}, -\bar{\theta} + \Lambda \mathbf{c} + \Lambda \mathbf{t})$.

### III. DERIVATION OF THE FORECASTING MODEL

In Section III-A we present the feasible region of a prototype building which can be representative of the ensemble of buildings. Subsequently, in Section III-B we provide the feasible region of an aggregate building which is built upon the prototype building by using the concept of homothet. Finally, in Section III-C we derive the forecasting model.

#### A. Building Prototype

We consider that the prototype building is the one representing the average behaviour of those in the pool. To do that, we have modeled the prototype building with a thermal model wherein all the physical and technical parameters $\mathcal{O}^p = \{R^p, C^p, \theta^p, \Delta, \eta^p, \theta_0^p, P^p\}$ are the mean of the parameters $\mathcal{O}^i$ for each building $i$. Once $\mathcal{O}^p$ is computed, we can obtain the set of model-related parameters $\Phi = \{c^p, P^p, \bar{p}^p, \theta^p, \bar{\theta}^p \}$ of the building prototype. We denote its feasible region as:

$$\mathbb{B}^p = \{ \mathbf{p}^p \in \mathbb{R}^{n_H} | \mathbf{F}^p \mathbf{p}^p \leq \mathbf{h}^p \}. \quad (6)$$

#### B. Aggregate Building Model

Similarly, we approximate the aggregate feasible region as another thermal building model:

$$\mathbb{B}^a = \{ \mathbf{p}^a \in \mathbb{R}^{n_H} | \mathbf{F}^a \mathbf{p}^a \leq \mathbf{h}^a \}. \quad (7)$$

However, the set of model-related parameters of the pool of buildings associated with (7), i.e., $\Phi^a = \{c^a, P^a, \bar{p}^a, \theta^a, \bar{\theta}^a \}$, are unknown. One possibility would be to infer all these parameters from observations of the aggregate power of the pool of buildings. However, this is most likely to be a lost cause (due to unobservability issues), lead to overfitting, and result in instability of the estimation algorithm.

To overcome such difficulty, we assume that the aggregate feasible region is a homothet of the prototype building, i.e., we can express $\mathbb{B}^a$ in terms of $\mathbb{B}^p$ as:

$$\mathbb{B}^a = \{ \mathbf{p}^a \in \mathbb{R}^{n_H} | \mathbf{F}^a \mathbf{p}^a = \beta \mathbf{p}^p + \tau, \forall \mathbf{p}^p \in \mathbb{B}^p \}, \quad (8)$$

where $\beta > 0$ is a scaling factor, and $\tau$ is a vector of translation factors. Expression (8) is the formal definition of a homothet for the aggregate power in $\mathbb{R}^{n_H}$.

By using the definition of homothet (8) and the prototype feasible region (6), we can recast the aggregate one as:

$$\mathbb{B}^a = \{ \mathbf{p}^a \in \mathbb{R}^{n_H} | \mathbf{F}^p \mathbf{p}^a \leq \mathbf{h}^a \beta + \mathbf{F}^p \tau \}. \quad (9)$$

By identifying the right-hand side of expression (9) with the right-hand side of expression (7), we can express the unknown...
vectors and matrices of the set $\Phi^a$ exclusively in terms of the homothetic parameters $\beta$ and $\tau$:

$$c^a = \beta c^b$$  \hspace{1cm} (10a)

$$\phi^a = \phi^b + \tau$$  \hspace{1cm} (10b)

$$p^a = \phi^a + \tau$$  \hspace{1cm} (10c)

$$t^a = \beta \theta^a b (1 - a_1)$$  \hspace{1cm} (10d)

$$\theta^a = \theta^b + \Lambda B \tau$$  \hspace{1cm} (10e)

$$\overline{\theta}^a = \overline{\theta}^b + \Lambda B \tau.$$  \hspace{1cm} (10f)

Lastly, the feasible region of the homothetic aggregate building can be formulated as follows:

$$p^a \leq \phi^a \leq \overline{p}^a$$  \hspace{1cm} (11a)

$$\theta^a \leq \Lambda B p^a + \Lambda c^a + \Lambda t^a \leq \overline{\theta}^a.$$  \hspace{1cm} (11b)

The set of model-related parameters $\Phi^a = \{c^a, \phi^a, \overline{p}^a, \overline{\theta}^a, \theta^a\}$ is parameterized in terms of the homothetic parameters $\beta$ and $\tau$, i.e., $\Phi^a(\beta, \tau)$ according to expressions (10). Therefore, the feasible region of the homothetic aggregate building (11) depends entirely on the homothetic parameters and the set of model-related parameters of the prototype building $\Phi^b$ (which are given), thus dramatically reducing the complexity of the model to be estimated and avoiding the undesirable overfitting effect. To the best of the authors’ knowledge, this is the first time in the literature that such a geometric approach is used to drastically simplify the task of forecasting the price-responsive aggregate power of a pool of buildings via inverse optimization, as explained below.

C. Forecasting Model

Let us assume that the feasible region of the ensemble of buildings is a homothetic representation of a prototype building and that the utility function of the pool is a step-wise price function with $n_B$ blocks. Under these assumptions and given the electricity prices, the forecasting model for each day $d$ can be mathematically expressed as:

$$\max_{p_{b,d}, \phi_d, s_d} \sum_{b \in B} (m^T_{b,d} - \lambda^T_d) p_{b,d} - e^s_T s_d^a$$  \hspace{1cm} (12a)

subject to:

$$\phi_d \leq \sum_{b \in B} \phi_{b,d} \leq \phi_d : (\phi_d, \overline{\phi}_d)$$  \hspace{1cm} (12b)

$$\theta_d - s_d^a \leq \sum_{b \in B} \Lambda B p_{b,d} + \Lambda (c_d + t_d) \leq \overline{\theta}_d + s_d^a : (\theta_d, \overline{\theta}_d)$$  \hspace{1cm} (12c)

$$0 \leq p_{b,d} \leq \tau_{b,d} : (\phi_{b,d}, \overline{\phi}_{b,d}), \forall b \in B$$  \hspace{1cm} (12d)

$$s_d^a \geq 0 : (\phi_d),$$  \hspace{1cm} (12e)

where $m_{b,d}, \lambda_{d,}$ and $e^s$ are the vectors of marginal utilities, electricity prices, and discomfort costs. The dual variables are shown in parentheses after a colon next to the corresponding constraints.

The objective function (12a) aims to maximize the welfare of the pool of buildings while minimizing the discomfort the users may experience. Constraints (12b)-(12c) are almost identical to the homothetic representation of the aggregate feasible region (11). Without loss of generality, we have incorporated some degree of flexibility into the forecasting model by: (i) modeling step-wise marginal utility functions to adequately learn the price-response of the pool of buildings, and (ii) including the slack variable in the temperature-related constraints (12c) to capture the infeasibilities that the modeling of building thermal dynamics may cause. Constraints (12d) impose lower and upper bounds on the aggregate power per block $b$, being $\tau_{b,d}$ the length of the power per block $b$ in day $d$. Finally, constraints (12e) set the non-negative character of slack variables.

As previously mentioned, the set $\Phi^a(\beta, \tau)$ is parameterized in terms of the homothetic parameters $\beta$ and $\tau$ according to expressions (10). Therefore, in this problem, the vector of marginal utilities $m_{b,d}$, as well as the homothetic parameters $\beta$ and $\tau$ are parameters to be estimated through IO, as explained in Section IV.

IV. INVERSE OPTIMIZATION METHODOLOGY

In this section, we thoroughly describe the proposed IO methodology to infer the parameters $m_{b,d}, \beta$, and $\tau$ of the forecasting model (12). First, we present the bilevel program and its transformation into a nonlinear single-level equivalent program. Then, we rewrite it into a parametric nonlinear problem that can be solved by commercial solvers. Later, we briefly describe a novel day-ahead two-step heuristic estimation approach, inspired in the one-step forecasting technique proposed in [5], that produces good estimates of the homothetic parameters $\beta$ and $\tau$. We then fix these estimates in the parametric nonlinear problem. Besides, the solution provided by the heuristics serves as a suitable initial point to the proposed nonlinear program. Finally, we list the steps of the proposed approach.

A. Bilevel Problem

Let us denote the vector of observed aggregate power in day $d$ as $p^a_d$. The goal of the bilevel problem is to minimize the mean absolute error of the aggregate power:

$$\min_{m_{b,d}, \phi_d, s_d, \beta, \tau, \nu, \rho} \sum_{d \in D} \sum_{b \in B} \left| p^a_d - p^a_d \right|$$  \hspace{1cm} (13a)

subject to:

$$m_{b,d} = \nu b + z_d \rho, \forall b \in B, d \in D$$  \hspace{1cm} (13b)

$$\nu_b \geq \nu_{b+1}, \forall b < n_B$$  \hspace{1cm} (13c)

Lower-Level Problem (12), $\forall d \in D$.  \hspace{1cm} (13d)

On the one hand, the upper-level problem (13a)-(13c) minimizes the absolute error of the estimated aggregate power of the pool with respect to the observed one, as given by (13a). Constraints (13b) impose linear regression functions, with $\nu_b$ and $\rho$ as the coefficients to be estimated, so that the marginal utilities are related to the regressors. Constraints (13c) set the marginal utilities to be monotonically non-increasing, as commonly done in electricity markets [14]. The lower-level problem (13d) is essentially the forecasting problem (12) for each day $d$. This lower level is solely parameterized in terms
of the marginal utilities $m_{b,d}$, as well as the homothetic parameters $\beta$ and $\tau$, and thus rendering the lower level as a linear program. Therefore, we can apply the Karush-Kuhn-Tucker necessary optimality conditions to the lower level in order to transform the original bilevel model (13) into the following nonlinear single-level equivalent:

$$\min_{\Xi^{NRP}} \mathbb{E}_{d \in D} \left( \sum_{b \in B} \left[ \left| p_{b,d}^o - p_{d}^o \right| \right]_1 \right)$$

subject to:

$$m_{b,d} = \nu_b + z_d \rho, \quad \forall b \in B, \forall d \in D$$

$$\nu_b \geq \nu_{b+1}, \quad \forall b < n_B$$

$$p_{d}^o \leq \sum_{b \in B} p_{b,d}^o \leq p_{d}^o, \quad \forall d \in D$$

$$\theta_d - s_d \leq \sum_{b \in B} \Lambda B p_{b,d}^o + \lambda (c_d^o + t_d^o) \leq \theta_d + s_d, \quad \forall d \in D$$

$$0 \leq p_{b,d}^o \leq \zeta_{b,d}, \quad \forall b \in B, \forall d \in D$$

$$s_d \geq 0, \quad \forall d \in D$$

$$\zeta_d = 0, \quad \forall d \in D$$

$$\phi_{b,d}^o, \varphi_{b,d}^o \geq 0, \quad \forall b \in B, \forall d \in D$$

$$\tau_d \left( \sum_{b \in B} p_{b,d}^o - p_{d}^o \right) = 0, \quad \forall d \in D$$

$$\kappa_d^T \left( \sum_{b \in B} \Lambda_{b} B_{b,d} p_{b,d}^o - \Lambda_d (c_d^o + t_d^o) \right) = 0, \forall d \in D$$

$$\zeta_{b,d}^o = s_{b,d}^o = 0, \quad \forall b \in B, \forall d \in D$$

$$\phi_{b,d}^o = 0, \quad \forall b \in B, \forall d \in D$$

where the set of decision variables is $\Xi^{NRP} = \left\{ m_{b,d}, p_{b,d}^o, \zeta_{b,d}, \kappa_d^T, \theta_d, \nu_b, \varphi_{b,d}^o, \tau_d, \beta, \tau \right\}$.

The rest of decision variables of the set $\Psi^o$ have been omitted because they are parameterized in terms of the homothetic parameters via expressions (10). For the sake of simplicity, we assume blocks of identical length, i.e., $\zeta_{b,d} = p_{b,d}^o / n_B$ for each block $b$ and day $d$.

Expressions (13a)–(13b) represent the upper-level problem while expressions (14a)–(14r) are the Karush-Kuhn-Tucker optimality conditions associated with the lower-level problem (12). The single-level equivalent (14) is characterized as a nonlinear program due to the products of continuous Lagrange multipliers and both upper- and lower-level decision variables in the complementarity slackness conditions (14r)–(14r). Typically, these constraints can be reformulated by way of binaries and big constants using the so-called Fortuny-Amat and McCarl linearization [15]. However, the mixed-integer linear counterpart of problem (14) becomes computationally intractable when moderately increasing the sample size. Besides, solving the nonlinear problem (14) as such may be also intractable since it is inherently ill-posed, as stated in [16]. To circumvent such an issue, we propose to recast the nonlinear single-level equivalent (14) into a regularized nonlinear program, as described in [16], [17].

### B. Regularized Nonlinear Problem

The single-level equivalent formulation (14) can be rewritten as follows:

$$\min_{\Xi^{NRP}} \mathbb{E}_{d \in D} \left( \sum_{b \in B} \left[ \left| p_{b,d}^o - p_{d}^o \right| \right]_1 \right)$$

subject to:

$$\sum_{b \in B} \left[ \left| p_{b,d}^o - p_{d}^o \right| \right]_1 \leq \sum_{d \in D} \left[ \left| p_{b,d}^o - p_{d}^o \right| \right]_1$$

$$\sum_{d \in D} \left[ \left| p_{b,d}^o - p_{d}^o \right| \right]_1 + \sum_{d \in D} \left[ \left| p_{b,d}^o - p_{d}^o \right| \right]_1$$

$$0 \leq p_{b,d}^o \leq \zeta_{b,d}, \quad \forall b \in B, \forall d \in D$$

$$s_d \geq 0, \quad \forall d \in D$$

$$\zeta_d = 0, \quad \forall d \in D$$

$$\phi_{b,d}^o, \varphi_{b,d}^o \geq 0, \quad \forall b \in B, \forall d \in D$$

$$\tau_d \left( \sum_{b \in B} p_{b,d}^o - p_{d}^o \right) = 0, \quad \forall d \in D$$

$$\kappa_d^T \left( \sum_{b \in B} \Lambda_{b} B_{b,d} p_{b,d}^o - \Lambda_d (c_d^o + t_d^o) \right) = 0, \forall d \in D$$

$$\zeta_{b,d}^o = s_{b,d}^o = 0, \quad \forall b \in B, \forall d \in D$$

$$\phi_{b,d}^o = 0, \quad \forall b \in B, \forall d \in D$$

where the set of decision variables is $\Xi^{NRP} = \Xi^{NBP}$. Note that the absolute value in (15a) can be reformulated as a set of linear expressions.

In problem (15), we basically relax the sum of all complementary slackness conditions (14r)–(14r) in (15c) by means of the parameter $\iota > 0$. When this parameter is sufficiently small, we can speed up the search of a locally optimal solution found by a nonlinear commercial solver. Hereinafter, we refer to problem (15) as $nmp$.

### C. Heuristic Estimation Approach

Another shortcoming of any nonlinear program is its sensitivity to the initial search point. In order to avoid meaningless local optima, we propose the use of efficient heuristics in which two convex programs are sequentially run to infer the marginal utilities $m_{b,d}$ and the homothetic parameters $\beta$ and $\tau$. The result of these heuristics, which are hereinafter referred to as $h2s$, can be utilized as a proxy of the IO problem (13), and thus can be used to yield more interpretable local optimal solutions from the regularized nonlinear problem (15).

The proposed heuristics are built upon the one put forward in [5], but the procedure has been substantially modified to account for the building thermal dynamics. The heuristics run as follows:

1. We first solve the feasibility problem. Its main goal is to estimate the parameters shaping the feasible region of the forward problem (12), by using the observed aggregate values of the pool of buildings. In this paper, the feasible region exclusively depends on the homothetic parameters...
D. Steps of the Proposed Approach

For the sake of clarity, we list the steps of the proposed forecasting technique, which we denote as $h2s + rnp$:

1. We first solve the two-step heuristic estimation process $h2s$ described in Section IV-C for a training set. This gives us a suitable proxy for the solution of the original bilevel problem (13). From this procedure, we obtain the marginal utilities $m_{h,d}$ and the corresponding estimates $v_h$ and $\rho$, as well as the homothetic parameters $\beta$ and $\tau$.

2. We then run the forecasting model (12) for the training set to compute the value of the in-sample aggregate power $P_{h,d}$, slack variable $s_h^b$, and the dual variables.

3. Afterwards, we run the regularized nonlinear program (15), denoted as $rnp$, with fixed homothetic parameters $\beta$ and $\tau$. In addition, we take the solution from the forecasting model (12) evaluated in the training set (previous step) as initialization. Here, we essentially re-optimize the marginal utility curves which may lead to an improved solution compared to the one provided by $h2s$.

4. Finally, the forecasting model (12) is built with the homothetic parameters $\beta$ and $\tau$ and the estimates $v_h$ and $\rho$ for the marginal utility curves.

\begin{table}[h]
\centering
\caption{Value of Power Block Upper Limit $\pi_{h,b}$}
\begin{tabular}{ccc}
\hline
$n_B = 1$ & $b \in B$ & $\pi_{h,b}^a = 0$ \\
$n_B > 1$ & $b = 1$ & $\pi_{h,n_B}^a = 0$ \\
& $b \in B \setminus \{1\}$ & $\pi_{h,n_B}^a = \frac{\pi_{h,n_B}^a - \pi_{h,n_B}^a}{n_B - 1}$ \\
\hline
\end{tabular}
\end{table}

\section{V. Comparison Methodologies}

We compare the forecasting capabilities of the proposed approach $h2s + rnp$ against five benchmarks: (i) the data-driven two-step estimation procedure $h2s$ outlined in Section IV-C; (ii) the regularized nonlinear problem $rnp$ formulated in Section IV-B with $\beta$ and $\tau$ free, without any initialization of the decision variables, and with $\iota = 0$; (iii) a simpler two-step estimation procedure taken from [5] and denoted as $s2s$; (iv) an AutoRegressive Integrated Moving Average Model with eXogenous (arimax) variables; and (v) a persistence model denoted as naive. The benefits of all models are compared by analyzing two error metrics: the mean absolute error (MAE) and the root mean square error (RMSE) on a test data set.

The forecasting problem associated with the heuristic $s2s$ is driven by the maximization of the welfare of the pool of buildings subject to solely the power bounds. In this benchmark, the indoor temperature bounds are ignored, thus overloading the effect of the building thermal dynamics. As similarly done in the proposed $h2s$, the marginal utilities and the power bounds are inferred by successively running two linear programs so that the RMSE is minimized in a validation data set. The interested reader is referred to [5] for a detailed description of the methodology.

The arimax model has been implemented in Python programming language [20] via the SARIMAX class of the package statsmodels. We have set the maximum number of iterations to 1000 and the stopping criterion is based on the estimator Akaike information criterion (AIC).

The forecast values of the aggregate power in day $d$ is equal to the observed values in the previous day $d-1$ for the naive model. The forecast error of this model is indicative of how hard predicting the demand of the pool of buildings is.

VI. CASE STUDY

We aim to learn the aggregate power of a population of 100 heterogeneous buildings. We first summarize the process to synthetically generate the data set. Then we present the input data for testing the forecast capabilities. Finally, we discuss the results obtained with the proposed approach $h2s + rnp$ and the benchmarks.

A. Data Generation for a Pool of Buildings

We assume that the consumption of each building $i$ for each day $d$ is driven by the following optimization problem:

$$\min_{p_h, s_h, \theta} \sum_{h \in \mathcal{H}} (p_h \lambda_h + g s_h)$$

$$\theta_h = a_1 \theta_{h-1} + (1 - a_1) \left[ \theta_{amb} - a_2 p_h \right], \forall h \in \mathcal{H}$$

$$-s_h + \theta_h \leq \theta_h + s_h, \quad \forall h \in \mathcal{H}$$

$$0 \leq p_h \leq \pi_h, \quad \forall h \in \mathcal{H}$$

$$s_h \geq 0, \quad \forall h \in \mathcal{H},$$

where $\theta$ represents a discomfort cost. Each building aims to minimize its electricity and discomfort costs, as in (16a), while satisfying the building thermal dynamics (16b), the temperature comfort bounds (16c), and the power bounds of...
the cooling device \[16d\]. Slack variables are declared non-negative in \[16e\].

The technical parameters \(\Omega^p\) for the prototype building are shown in Table II. As done in \[15\], the model parameters \(\Omega^i\) for each building \(i\) of the pool are assumed to be uniformly distributed based on a factor \(h\) modeling the degree of heterogeneity. For instance, the samples for the thermal capacitance \(C^i\) are drawn from a uniform distribution in the interval \([1 - h]C^p, (1 + h)C^p\], where \(C^p\) is the thermal capacitance of the prototype building. The discretization step \(\delta\) is assumed to be one hour and the discomfort cost \(g\) is set to 0.01 \(\text{€/°C}\) for all buildings. Ambient temperature, electricity prices, and the building technical parameters \(\Omega^i\) are given in \[18\], for the sake of reproducibility.

### B. Input Data for Testing the Forecasting Models

We run simulations for 1872 hours (78 days) by using model \[16\] for two different values of the heterogeneity factor: (i) \(h = 0.1\) (low heterogeneity among buildings), and (ii) \(h = 0.75\) (high heterogeneity among buildings). To avoid undesirable border effects, we disregard the results from the first day of the simulation. The sizes for the training, validation, and test sets are 35, 35, and 7 days, respectively, in chronological order. For each case, reference \[18\] includes the aggregate power and the initial indoor air temperature per day.

Table III summarizes some statistics on the aggregate power.

| \(h = 0.1\) | \(h = 0.75\) |
|---|---|
| Maximum (kW) | 541.0 | 218.4 |
| Mean (kW) | 64.0 | 42.0 |
| Total (MW) | 118.2 | 77.6 |
| # hours without consumption (%) | 61.8 | 0.0 |

C. Results

We analyze the impact of the degree of heterogeneity of the pool of buildings on the forecasting capabilities of the proposed technique. Besides, for the models \(h2s + rnp\), \(h2s\), \(rnp\), and \(s2s\), we further study the behavior of the models when considering either (i) a number of power blocks \(n_B = 1\), and (ii) \(n_B = 6\). Table IV provides the forecast error metrics, namely RMSE and MAE, for all models outlined in Section V and the aforementioned cases. In this table, we highlight the best results in bold.

First, we discuss the results for an heterogeneity factor \(h = 0.1\) when considering a single power block, i.e. \(n_B = 1\). In this setup, the proposed method \(h2s + rnp\) leads to the best forecasting performance in terms of RMSE and MAE since they are reduced by 39.9% and 41.7% compared to the naive model. The proposed technique \(h2s + rnp\) enhances the MAE in a 2.0% with respect to the model \(h2s\) while keeping the RMSE at almost the same value. Notwithstanding, solving the model \(rnp\) by itself, i.e. without fixing the homothetic parameters, without using the initialization given by the heuristic \(h2s\), and with \(\epsilon = 0\), raises the RMSE and MAE by 54.9% and 44.8%, in that order, regarding the ones obtained with the proposed model \(h2s + rnp\).

The nonlinear model \(rnp\) converges to the local optimal solution with \(\beta = 0\), which seems to be an attractive solution due to the nature of this mathematical problem. The reason behind this outcome relies on the definition of homothet \[8\] and the subsequent derivations of the model-related parameters \[10\]. \(\beta = 0\) implies a constant objective function in the lower-level problem \[13\] and a feasible region that boils down to the singleton \(p_s^d = \tau, \forall d \in D\), according to the definitions mentioned above. Therefore, the upper level \[13\] basically seeks the vector \(\tau\) minimizing the MAE over the training data set. Finally, both models \(s2s\) and arimax attain higher forecasting errors than the proposed technique \(h2s + rnp\), i.e. RMSE increases by 25.4% and 51.6%, in that order, whereas the respective increase in MAE is 66.0% and 105.5%. The reason of those poor forecasts is because both models disregard the effect of the building thermal dynamics in the forecasting process.

For the case of \(h = 0.1\), when increasing the number of power blocks to 6, the proposed method \(h2s + rnp\) refines the solution achieved with only one block, i.e., RMSE and MAE decreases by 2.8% and 0.4%, in that order. This is an indication that there is a small degree of sensitiveness of the power to the price, which is captured by means of the optimality problem of the heuristic \(h2s\). In this setup, the model \(h2s + rnp\) has lowered the RMSE and MAE by 1.0% and 1.7%, respectively, over the solution given by the heuristic \(h2s\).

Table IV also shows the results when there is a high heterogeneity among buildings, i.e., \(h = 0.75\). In this situation, there is a higher power-price sensitivity and the aggregate power becomes smoother along the time. For this reason, the naive model provides a better accuracy compared to the results with a low heterogeneity factor, i.e. RMSE = 36.9 and MAE = 24.2. In terms of RMSE, the proposed technique \(h2s + rnp\) exhibits the best forecasting performance with \(n_B = 1,\)

### Table II: Technical Parameters \(\Omega^p\) for the Prototype Building

| \(C\) (kWh/°C) | 10 | \(\theta\) | 2.5 | \(\Delta\) (°C) | 1 |
|---|---|---|---|---|---|
| \(R\) (°C/kWh) | 2 | \(\theta^r\) (°C) | 20 | \(b_i\) (°C) | 22.5 |
| \(P\) (kW) | 5.4 |  |

### Table III: Statistics on the Aggregate Power

| \(h = 0.1\) | \(h = 0.75\) |
|---|---|
| Maximum (kW) | 541.0 | 218.4 |
| Mean (kW) | 64.0 | 42.0 |
| Total (MW) | 118.2 | 77.6 |
| # hours without consumption (%) | 61.8 | 0.0 |
TABLE IV

Error Metrics – Comparison of Models

| Model | $h = 0.1$ | $h = 0.75$ |
|-------|-----------|-----------|
|       | $n_B = 1$ | $n_B = 6$ | $n_B = 1$ | $n_B = 6$ |
| $RMSE$ | MAE | $RMSE$ | MAE | $RMSE$ | MAE | $RMSE$ | MAE |
| $h2s + rnp$ | 106.7 | 52.7 | 103.7 | 52.5 | 31.3 | 22.5 | 25.3 | 16.9 |
| $h2s$ | 106.6 | 53.8 | 104.8 | 53.4 | 32.5 | 23.8 | 25.2 | 17.5 |
| $rnp$ | 165.3 | 76.3 | 165.3 | 76.3 | 35.9 | 22.9 | - | - |
| $s2s$ | 132.8 | 87.5 | 133.0 | 88.9 | 38.4 | 24.0 | 30.3 | 26.0 |
| $arimax$ | 161.8 | 108.3 | 161.8 | 108.3 | 31.6 | 23.0 | 31.6 | 23.0 |
| $naive$ | 177.5 | 90.4 | 177.5 | 90.4 | 36.9 | 24.2 | 36.9 | 24.2 |

which results in a 15.2%, 18.5%, 12.8%, 3.7%, and 0.9% of improvement over the naive error, $s2s$, $rnp$, $h2s$, and $arimax$ models, in that order. However, this improvement over the naive model (15.2% in RMSE) is substantially lower than when the buildings are more alike, wherein there is a reduction of 39.9% in RMSE. In addition, for the case of high heterogeneity, the forecasting performance of the proposed technique considerably enhances the error metrics when increasing the number of blocks to $n_B = 6$. Specifically, there is a reduction of 19.2% in RMSE and 24.9% in MAE over the solution with only one block whereas this reduction is just 2.8% and 0.4%, respectively, for $h = 0.1$. A similar observation can be made for $s2s$, which is based on similar heuristics. Notwithstanding, the forecasting capabilities of the model $s2s$ are by far worse than those from the proposed technique due to the fact that the former ignores the building thermal dynamics in the forecasting process. Finally, the model $rnp$ is unable to find a solution and the solver CONOPT returns an evaluation error. This is probably caused because the model $rnp$ with $t = 0$ is inherently ill-posed, as pointed out in [16].

VII. CONCLUSION

This paper has proposed a novel day-ahead forecasting technique for an aggregation of smart buildings equipped with thermostatically-controlled loads. From a modeling perspective, the aggregate power of the pool of buildings is represented by using a geometric approach, i.e., its price-response is characterized by a set of marginal utility curves and a homothet of a prototype building. This intuitive representation of the aggregate allows us to account for the building thermal dynamics while drastically reducing the number of parameters to be estimated. Hence, the computational complexity of the estimation algorithm is decreased, thus avoiding the undesirable overfitting effect. From a methodological perspective, inverse optimization is applied to infer the marginal utilities and the homothetic parameters by means of bilevel programming and an efficient heuristic estimation procedure. We can conclude that (i) accounting for the building thermal dynamics in the forecasting technique improves the forecasting error by 20–40% compared to existing and persistence methodologies when the buildings are more alike, and that (ii) the use of an increasing number of blocks for the marginal utilities in the forecasting process substantially improves the accuracy of the proposed forecasting technique when the heterogeneity among buildings is high. Future work will be devoted to further improving the accuracy by increasing the number of geometric parameters (e.g. a rotation of the homothet). Besides, we will explore the extension of the geometric parameters to be regressor-dependent.

REFERENCES

[1] P. Pinson, H. Madsen, et al., “Benefits and challenges of electrical demand response: A critical review,” Renewable and Sustainable Energy Reviews, vol. 39, pp. 686–699, 2014.
[2] M. Shahidehpour, H. Yamin, and Z. Li, Market operations in electric power systems: forecasting, scheduling, and risk management. John Wiley & Sons, 2003.
[3] O. Corradi, H. Ochsenfeld, H. Madsen, and P. Pinson, “Controlling electricity consumption by forecasting its response to varying prices,” IEEE Transactions on Power Systems, vol. 28, no. 1, pp. 421–429, 2012.
[4] M. Zhang, Quan, S. Caixin, L. Shaoan, L. Yuming, and S. Yang, “RBF neural network and ANFIS-based short-term load forecasting approach in real-time price environment,” IEEE Transactions on power systems, vol. 23, no. 3, pp. 853–858, 2008.
[5] J. Saez-Gallego and J. M. Morales, “Short-term forecasting of price-responsive loads using inverse optimization,” IEEE Transactions on Smart Grid, vol. 9, no. 5, pp. 4805–4814, 2018.
[6] A. Aswani, Z.-J. Shen, and A. Siddiqi, “Inverse optimization with noisy data,” Operations Research, vol. 66, no. 3, pp. 870–892, 2018.
[7] P. M. Esfahani, S. Shafieezadeh-Abadeh, A. G. Hanususanto, and D. Kuhn, “Data-driven inverse optimization with imperfect information,” Math. Progr., vol. 167, no. 1, pp. 191–234, 2018.
[8] P. Bennett, P. V. Gupta, and I. C. Paschalidis, “Data-driven estimation in equilibrium using inverse optimization,” Math. Progr., vol. 153, no. 2, pp. 595–633, 2015.
[9] T. C. Chan and N. Kaw, “Inverse optimization for the recovery of constraint parameters,” European Journal of Operational Research, vol. 282, no. 2, pp. 415–427, 2020.
[10] J. Saez-Gallego, J. M. Morales, M. Zugno, and H. Madsen, “A data-driven bidding model for a cluster of price-responsive consumers of electricity,” IEEE Transactions on Power Systems, vol. 31, no. 6, pp. 5001–5011, 2016.
[11] T. Lu, Z. Wang, J. Wang, Q. Ai, and C. Wang, “A data-driven stackelberg market strategy for demand response-enabled distribution systems,” IEEE Transactions on Smart Grid, vol. 10, no. 3, pp. 2345–2357, 2018.
[12] J. E. Contreras-Ocaña, M. A. Ortiga-Vazquez, D. Kirsch, and B. Zhang, “Tractable and robust modeling of building flexibility using coarse data,” IEEE Transactions on Power Systems, vol. 33, no. 5, pp. 5456–5468, 2018.
[13] L. Zhao, W. Zhang, H. Hao, and K. Kalsi, “A geometric approach to aggregate flexibility modeling of thermostatically-controlled loads,” IEEE Transactions on Power Systems, vol. 32, no. 6, pp. 4721–4731, 2017.
[14] OMIE, “Spanish Market Operator.” [Online]. Available: http://www.omie.es/, 2019.
[15] J. Fortuny-Amat and B. McCarl, “A representation and economic interpretation of a two-level programming problem,” Journal of the operational Research Society, vol. 32, no. 9, pp. 783–792, 1981.
[16] S. Scholtes, “Convergence properties of a regularization scheme for mathematical programs with complementarity constraints,” SIAM Journal on Optimization, vol. 11, no. 4, pp. 918–936, 2001.
[17] S. Pineda, H. Bylling, and J. Morales, “Efficiently solving linear bilevel programming problems using off-the-shelf optimization software,” Optimization and Engineering, vol. 19, no. 1, pp. 187–211, 2018.
[18] R. Fernández-Blanco, J. M. Morales, and S. Pineda, “Forecasting the price-response of a pool of buildings via homothetic inverse optimization — Auxiliary data.” [Online]. Available: shorturl.at/RU23, 2020.
[19] D. G. Luenberger, Y. Ye, et al., Linear and nonlinear programming, vol. 2. Springer, 1984.
[20] Python Software Foundation. Python Language Reference, version 3.7.3. [Online]. Available: https://www.python.org/.
[21] “IBM ILOG CPLEX Optimisation Studio.” [Online]. Available: https://www.ibm.com/analytics/cplex-optimizer.
[22] W. E. Hart, J. P. Watson, and D. L. Woodruff, “Pyomo: Modeling and solving mathematical programs in Python,” Mathematical Programming Computation, vol. 3, no. 3, pp. 219–260, 2011.
[23] “ARKI Consulting & Development AS.” [Online]. Available: http://www.conopt.com/.
[24] J. Czyzyk, M. P. Mesnier, and J. J. Moré, “The NEOS server,” IEEE Journal of Computational Science and Engineering, vol. 5, no. 3, pp. 68–75, 1998.