Scalable and Robust Demand Response With Mixed-Integer Constraints

Seung-Jun Kim and Georgios B. Giannakis

***Abstract*—A demand response (DR) problem is considered entailing a set of devices/subscribers, whose operating conditions are modeled using mixed-integer constraints. Device operational periods and power consumption levels are optimized in response to dynamic pricing information to balance user satisfaction and energy cost. Renewable energy resources and energy storage systems are also incorporated. Since DR becomes more effective as the number of participants grows, scalability is ensured through a parallel distributed algorithm, in which a DR coordinator and DR subscribers solve individual subproblems, guided by certain coordination signals. As the problem scales, the recovered solution becomes near-optimal. Robustness to random variations in electricity price and renewable generation is effected through robust optimization techniques. Real-time extension is also discussed. Numerical tests validate the proposed approach.**

***Index Terms*—Lagrange relaxation, mixed-integer programs, parallel and distributed algorithms, real-time demand response, robust optimization.**

A	Number of DR-capable devices (loads).
$\mathcal{A}$	Set of loads.
$\mathcal{A}_1$	Set of inelastic loads.
$\mathcal{A}_2$	Set of elastic and interruptible loads.
$\mathcal{A}_3$	Set of elastic and non-interruptible loads.
b^t	Total energy stored in energy storage devices at the end of the tth time slot.
$B_{\max}$	Capacity of energy storage devices.
$C^t(g^t)$	Electricity price at time t for power level g^t.
c_n^t	Slope of the nth segment of piecewise linear convex function $C^t(\cdot)$.
$\bar{c}_n^t$	Nominal value of c_n^t.
$\tilde{c}_n^t$	Max. deviation of c_n^t from the nominal value.
$E_a^{\rm tot}$	Total energy to be expended by elastic load a.
$E_{{\rm tot},a}^{\tau}$	Energy remaining to be expended at the beginning of the τth time slot.
E_a^t	Total energy expended by device a by the end of the tth time slot.
g^t	Power drawn from the grid at time t.
$G_{\max}$	Max. power that can be drawn from the grid.
l^t	Total load at time t.
p_a^t	Power consumed by device a at time t.
$\mathbf{p}_a$	Power consumption profile of device a.
$\mathcal{P}_a$	Set of admissible power profiles of device a.
$\mathcal{P}_a^{\tau}$	Set of admissible power profiles of device a at time τ under real-time DR.
$P_{\max},a$	Max. power consumed by device a when ON.
$P_{\min,a}$	Min. power consumed by device a when ON.
T	Length of DR time horizon.
$\mathcal{T}$	DR time horizon.
u_a^t	ON/OFF status of device a.
$U_a(\mathbf{p}_a)$	Utility gained from device a using power $\mathbf{p}_a$.
r^t	Energy stored into or drawn from storage devices at time t.
$R_{\max}$	Max. power at which energy can be stored
$R_{\min}$	Max. power at which energy can be drawn.
w^t	Power from renewable sources at time t.
$\bar{w}^t$	Nominal value of w^t
$\tilde{w}^t$	Max. deviation of w^t from the nominal value.
α_a	Time at or after which device a may be turned ON.
β_a	Time after which device a must be turned OFF.
Γ_C	Tuning parameter for robustness against electricity price uncertainty.
Γ_C^{τ}	Counterpart for Γ_C for real-time DR at time τ.
Γ_W	Tuning parameter for robustness against renewable generation uncertainty.
Γ_W^{τ}	Counterpart for Γ_W for real-time DR at time τ.
$\boldsymbol{\lambda}$	Lagrange multiplier vector.
θ_n^t	Breakpoints of piecewise linear convex $C^t(\cdot)$.

Manuscript received October 02, 2012; revised February 06, 2013; accepted April 03, 2013. Date of publication April 30, 2013; date of current version November 25, 2013. Part of this work was presented at the 4th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), San Juan, Puerto Rico, Dec. 2011. This work was supported in part by the Institute of Renewable Energy and the Environment (IREE) Grant RL-0010-13, University of Minnesota.

The authors are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: seungjunv@umn.edu; georgios@umn.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSG.2013.2257893

I. Introduction

DEMAND response (DR) is a key component of the smart grid, which allows both consumers and utilities to benefit through intelligent resource scheduling. By adopting electricity prices that vary depending on the time of use and the load level, power system operators can elicit desirable usage patterns from

consumers, saving costs related to generation, transmission and storage, and lowering customers' bills [1].

DR becomes more effective when the number of devices participating in the DR program grows large, as it will be more likely to have adjustable loads available when the need arises to reduce or increase the demand. DR aggregators capitalize on this by mediating power utilities and consumers, scheduling subscribers' power consumption to minimize cost and provide ancillary services. Thus, it is essential to solve large-scale DR problems efficiently, which favors parallelized and distributed solution approaches.

Another important issue is privacy. Under a centralized architecture, DR participants submit private data related to their desired power consumption charateristics to a central entity. A distributed solution can mitigate such concerns through exchanging only surrogate coordination signals [2].

To best capture the gains from DR and to ensure customer satisfaction, diverse usage profiles and operational constraints of the devices must be taken into account. Three salient types of devices are considered in this work. The first type is the *inelastic* load, which cannot be deferred to a later time, but its power consumption may be traded off with the user's degree of satisfaction. An example is the HVAC system. The second type of load is *elastic* and *interruptible*, which means that the power consumption may be shifted in time as long as a specified amount of total energy is expended, and its operation is allowed to be suspended in the middle. Examples include electric vehicles (EVs) charging and pool pumps. The third class is an *elastic* yet *non-interruptible* kind, which, once switched on, cannot be turned off until the task is completed, often following prespecified power usage patterns over time. Washing machines or dish washers in the residential setting fall into this class, as well as ample examples in the industrial context. While certain restricted combinations of these requirements lead to convex optimization problems [2], more general and descriptive constraints necessitate mixed-integer nonconvex formulations [3]–[6].

DR algorithms must cope with various uncertain parameters that constitute their input. The algorithms use electricity prices announced or forecast prior to the planning horizon, but the prices are subject to real-time amendment or forecasting errors. When renewable energy resources are utilized, it is critical to take into account the volatility of such resources, as they cannot be dispatched at will, and the reserve may be costly. Thus, it is prudent to design DR controls with robustness to ensure cost-effective operation even under unfavorable conditions. A robust optimization framework, successfully applied in DR contexts for linear pricing in [7], is employed here for piecewise linear convex pricing as well as renewable resources. Moreover, it is beneficial to allow real-time adaptation of DR schedules, as part of the uncertainties will have been eliminated by the time of use.

The goal of the present work is to pursue both optimality and scalability in large-scale DR mixed-integer formulations, with robustness to uncertainties in price prediction and renewable generation. Key to this feat is the Lagrange relaxation approach, which decomposes the overall problem to multiple subproblems, each of which can be solved in a decentralized and parallelized fashion. As the number of participating DR devices/subscribers increases, the solution recovered from Lagrange relaxation nears the optimum. The approach can also be extended to real-time DR. The Lagrange relaxation approach was employed for a convex DR formulation in [8]. This is the first attempt to solve *nonconvex* DR problems in the dual domain for scalability and near-optimality while incorporating robustness to major sources of uncertainties.

Prior works focused on a subset of task attributes to obtain tractable formulations [7], [9]. The difficulty of accommodating non-interruptible yet deferrable tasks was recognized in [10]. When applied to large-scale DR problems, DR formulations may require excessive computational complexity. Thus, suboptimal approximate solutions have often been advocated [4], [11]. In the absence of coupling constraints among tasks, justified in a simple residential setup, scheduling of non-interruptible tasks under price uncertainty is tractable [12].

The rest of the paper is organized as follows. Section II describes the system model and presents the basic DR formulation. Section III incorporates robustness against the uncertainties in electricity prices and renewable resources. Scalable solutions based on Lagrange relaxation are derived in Section IV. The real-time adaptation of DR schedules is discussed in Section V. Numerical examples are provided in Section VI, and conclusions are offered in Section VII.

Notational conventions: Superscripts ^{t} and $^\tau$ refer to time in a DR horizon, subscript $_a$ indicates DR devices, and subscript $_n$ different segments of piecewise linear functions. Superscript ^{k} is an iteration index, and ^{j} indexes individual elements in a bundle in the context of the bundle method.

II. Problem Formulation

To simplify exposition, a DR problem for multiple subscribers, each with just one DR-capable device (load) is considered. However, the proposed framework is general and capable of handling multiple devices per subscriber, as will be briefly explained later. First, a model for various device classes is established.

A. Device Model

A set $\mathcal{A} := \{1, 2, \ldots, A\}$ of devices is scheduled over a time horizon of T units, in response to the electricity price announced (or, forecast possibly with some error) [7], [9]. Let p_a^t denote the power consumption of device $a \in \mathcal{A}$ at time $t \in \mathcal{T} := \{1, 2, \ldots, T\}$. Collect the temporal power usage profile to a vector $\mathbf{p}_a := \{p_a^t\}_{t\in\mathcal{T}}$. The device constraints are captured by a set of admissible power usage profiles as in $\mathbf{p}_a \in \mathcal{P}_a$. It is assumed that power $\mathbf{p}_a$ consumed by device $a \in \mathcal{A}$ imparts some sort of satisfaction or *utility* to the user, which can be quantified by a function $U_a(\mathbf{p}_a)$.

Depending on the scheduling requirements, three types of loads are considered. The set of *inelastic* loads is denoted as $\mathcal{A}_1$. This class of loads cannot be deferred, although their power consumption may be traded off with the utility that users enjoy. Consider a simple power constraint that device $a \in \mathcal{A}_1$ consumes power between $P_{\min,a}$ and $P_{\max,a}$ when turned on. Suppose also that the device needs to be turned on only between times α_a and β_a. Introducing a binary variable $u_a^t \in \{1, 0\}$ that

models the ON/OFF status of device a for $t \in \mathcal{T}$, one can capture the aforementioned constraints by

$$P_{\min,a} u_a^t \leq p_a^t \leq P_{\max,a} u_a^t \quad (1)$$
$$u_a^t \in \{1,0\} \text{ if } t \in [\alpha_a, \beta_a] \quad (2)$$
$$u_a^t = 0 \text{ if } t < \alpha_a \text{or } t > \beta_a. \quad (3)$$

The second and third types of loads (denoted by $\mathcal{A}_2$ and $\mathcal{A}_3$, respectively) are *elastic* (deferrable) loads, which may be shifted in time as long as a certain amount E_a^{tot} of total energy is expended. Thus, in addition to (1)–(3), one imposes

$$\sum_{t\in\mathcal{T}} p_a^t \geq E_a^{\text{tot}}. \quad (4)$$

Different from the *interruptible* devices denoted by $\mathcal{A}_2$, the *non-interruptible* devices $a \in \mathcal{A}_3$ are additionally constrained that they cannot be turned off once switched on, until the desired total energy has been expended. Specifically,

$$u_a^t = 1 \text{ if } E_a^{t-1} := \sum_{\bar{t}=1}^{t-1} p_a^{\bar{t}} < E_a^{\text{tot}} \text{ and } u_a^{t-1} = 1,\ t \in \mathcal{T} \quad (5)$$

where u_a^0 is the given constant signifying the ON/OFF status of device a at the beginning of the first time slot. Note that (5) can be encoded into a mixed-integer linear constraint given as [13, Ch. 9]

$$E_a^{t-1} - (E_a^{\text{tot}} + 1 - \epsilon_{\text{m}}) u_a^{t-1} + E_a^{\text{tot}} u_a^t + 1 - \epsilon_{\text{m}} \geq 0 \quad (6)$$

where $\epsilon_{\text{m}} > 0$ is a small positive number (usually the smallest representable by the machine precision). To make sure that the elastic loads $a \in \mathcal{A}_2 \cup \mathcal{A}_3$ stay turned off once done, one can additionally require

$$u_a^t = 0 \text{ if } E_a^{t-1} \geq E_a^{\text{tot}} \quad (7)$$

which is equivalent to $E_a^{t-1} + (P_{\max,a} + \epsilon_{\text{m}}) u_a^t - P_{\max,a} - E_a^{\text{tot}} \leq 0$. In summary, the feasible power sets are defined as

$$\mathcal{P}_a := \begin{cases} \{\mathbf{p}_a : (1)-(3)\}, & a \in \mathcal{A}_1 \\ \{\mathbf{p}_a : (1)-(4) \text{ and } (7)\}, & a \in \mathcal{A}_2 \\ \{\mathbf{p}_a : (1)-(5) \text{ and } (7)\}, & a \in \mathcal{A}_3 \end{cases}. \quad (8)$$

It can be seen that if $P_{\min,a} = 0$, $\mathcal{P}_a$ becomes convex, and the use of integer variables is not necessary. In particular, the distinction between interruptible and non-interruptible classes vanishes. However, it is often the case in practice that nonzero minimum power must be consumed when a device is on. This requirement renders the constraint sets nonconvex. It will be also useful later to note that $\mathcal{P}_a$ becomes convex if the integer variables $\mathbf{u}_a := \{u_a^t\}_{t\in\mathcal{T}}$ are fixed.

To model the user utilities of various device classes, functions with simple structures are considered. For device $a \in \mathcal{A}_1$, let $U_a^t(p_a^t)$ denote the per-time-slot utility that depends only on the power consumption of the current time slot p_a^t. Then, the overall utility is defined as

$$U_a(\mathbf{p}_a) := \sum_{t\in\mathcal{T}} U_a^t(p_a^t),\ a \in \mathcal{A}_1. \quad (9)$$

As for the utilities of elastic devices, consider

$$U_a(\mathbf{p}_a) := -\sum_{t\in\mathcal{T}} D_a^t \left(E_a^{\text{tot}} - E_a^{t-1}\right),\ a \in \mathcal{A}_2 \cup \mathcal{A}_3 \quad (10)$$

where convex $D_a^t(E)$ models the "disutility" due to the energy of amount E remaining to be scheduled at the beginning of time slot t.

Remark 1: Typically, $D_a^t(\cdot)$ is non-decreasing with $D_a^t(E) = 0$ for $E \leq 0$, promoting early completion of elastic tasks, and penalizing deferred loads. One can also penalize "load advancement," by having non-increasing $D_a^t(\cdot)$ with $D_a^t(E) = 0$ for $E \geq E_a^{\text{tot}}$.

B. Electricity Price

Function $C^t(g^t)$ models the electricity price at time t for power g^t delivered from the grid. The cost function captures time and load-dependent electricity cost, and can be used to incentivize peak demand reduction. A convex $C^t(\cdot)$ is used to model the progressive price increase for higher load. A piecewise linear convex function is often employed for computational merits [9]. Note that in the special case where $C^t(\cdot)$ is linear and the load is served entirely on the power purchased from the grid, the overall cost $C^t\left(\sum_{a\in\mathcal{A}} p_a^t\right)$ can be re-expressed as the sum of the cost due to the individual devices. The power drawn from the grid must obey $g^t \in \mathcal{G} := [0, G_{\max}]$ for $t \in \mathcal{T}$, where $G_{\max}$ represents the cap on the total deliverable power, e.g., due to the capacity of transmission/distribution lines.

C. Renewable Generation and Storage

Renewable energy sources such as wind turbines or photovoltaic (PV) generators are incorporated in this section, along with energy storage devices such as batteries, flywheels or hydrogen storage. The amount of power w^t generated from the renewable sources can be forecast over time horizon $\mathcal{T}$. To model energy storage and retrieval, let b^t be the total energy stored at the end of the tth time slot. Let r^t represent the energy stored if $r^t > 0$, and $-r^t$ the energy drawn from the storage if $r^t < 0$, respectively. Then,

$$b^t = b^{t-1} + r^t,\ t \in \mathcal{T} \quad (11)$$

holds, where ideal energy conversion efficiency has been assumed for simplicity. Energy storage devices may have ramp constraints modeled as $r^t \in \mathcal{R} := [-R_{\min}, R_{\max}]$, where $R_{\min}$ and $R_{\max}$ are the maximum power that can be drawn from or charged to the storage per time interval, respectively. Also, due to the capacity $B_{\max}$ of the energy storage, $b^t \in \mathcal{B} := [0, B_{\max}]$ holds for each $t \in \mathcal{T}$.

D. Optimization Problem

The DR problem can now be formulated as

$$\min_{\substack{\mathbf{p}_a\in\mathcal{P}_a, b^t\in\mathcal{B},\\ g^t\in\mathcal{G}, r^t\in\mathcal{R}, l^t}} \sum_{t\in\mathcal{T}} C^t(g^t) - \sum_{a\in\mathcal{A}} U_a(\mathbf{p}_a) \tag{12a}$$

$$\text{subject to } l^t = \sum_{a\in\mathcal{A}} p_a^t,\ t\in\mathcal{T} \tag{12b}$$

$$0 \le r^t + l^t \le g^t + w^t,\ t\in\mathcal{T} \tag{12c}$$

$$b^t = b^{t-1} + r^t,\ t\in\mathcal{T}. \tag{12d}$$

Here, l^t represents the total load at time t, which is the sum of power consumption of all devices as expressed in (12b). The left inequality in (12c) constrains the power drawn from the storage not to exceed the actual load. The right inequality in (12c) states that the total load is served by the power drawn from the grid, the renewable source, and the storage. The initial state of the energy storage device b^0 is a given constant.

The objective in (12a) pursues balancing the energy cost and the subscribers' satisfaction. Relative importance of the two terms can be represented by weights, which can be absorbed into the definition of the utility functions $U_a(\cdot)$. Fairness among the subscribers' power consumption can also be handled by setting the weights appropriately. Such a strategy is often adopted for fair scheduling in wireless networks [14].

In the case of having multiple DR-capable devices per subscriber, set $\mathcal{A}$ can be interpreted as the set of *subscribers*, and $\mathbf{p}_a$ captures the total power usage accounting for all devices in the premises of subscriber $a\in\mathcal{A}$. Utility $U_a(\mathbf{p}_a)$ should also be viewed as the aggregate utility of subscriber a's devices. The feasible power set $\mathcal{P}_a$ can also be defined similarly, respecting the characteristics of individual devices.

III. Robust Formulation

The optimization problem (12) assumes that the electricity price is announced in advance, and does not change at the time of use. Also, the amount of renewable energy available per time instant is assumed to be perfectly known, or predicted accurately. In practice, the price may change in real time, reflecting the uncertainty of the load and the contingencies of grid operation. Renewable energy resources are random in nature, leading to forecasting errors.

To cope with uncertainties, a robust optimization approach is taken in this section, which essentially tackles the worst case. That is, upon presuming bounded uncertainty, the solution that remains feasible in all cases at a minimal sacrifice of optimality is pursued [15]. An important issue in robust optimization is to strike a reasonable trade-off between robustness and optimality [16]. If one pursues robustness against too large an uncertainty region, the corresponding solution may become unacceptably conservative and inefficient.

To mitigate this issue, techniques that allow only up to a preset number of uncertain parameters to deviate from the nominal values have been proposed [16], [17]. By adjusting the number of deviating parameters, the desired level of protection can be achieved. Moreover, the complexity of the robust version is on par with the original problem. The technique was applied in the DR context to price uncertainty in [7] for linear pricing. This is extended here to account for piecewise linear convex prices, as well as renewable resources.

A limitation of the approaches in [16] and [15] is that they require the worst-case protection adopted *per* constraint, and do not truly correspond to the min-max approach that minimizes the cost against the maximum degradation due to the deviation of uncertain parameters present *across* multiple constraints [18]. The latter formulation is often very difficult to solve. Such an issue emerges when incorporating renewable energy resources, but is mitigated here using a simple heuristic considered in a different application also by [19].

A. Robustness Against Price Uncertainty

Consider a piecewise linear convex function $C^t(g)$ defined for $g\ge 0$ with N_C^t segments. Suppose that the slopes of the segments satisfy $c_1^t \le c_2^t \le \cdots \le c_{N_C^t}^t$, and the breakpoints $0<\theta_1^t<\cdots<\theta_{N_C^t-1}^t$. Then, (12a) can be re-written as [20]

$$\min_{\substack{\mathbf{p}_a\in\mathcal{P}_a, b^t\in\mathcal{B},\\ g^t\in\mathcal{G}, r^t\in\mathcal{R}, l^t, x_n^t\ge 0}} \sum_{t\in\mathcal{T}}\sum_{n=1}^{N_C^t} c_n^t x_n^t - \sum_{a\in\mathcal{A}} U_a(\mathbf{p}_a) \tag{13}$$

with the following additional constraints:

$$x_1^t \le \theta_1^t,\ t\in\mathcal{T} \tag{14}$$

$$x_n^t \le \theta_n^t - \theta_{n-1}^t,\ n=2,\ldots,N_C^t-1,\ t\in\mathcal{T} \tag{15}$$

$$\sum_{n=1}^{N_C^t} x_n^t = g^t,\ t\in\mathcal{T} \tag{16}$$

where auxiliary x_n^t, for $n=1,\ldots,N_C^t$, signifies the portion of g^t that lies in $(\theta_{n-1}^t,\theta_n^t]$ (with $\theta_0^t:=0$ and $\theta_{N_C^t}^t:=G_{\max}$).

To model price uncertainty, assume that the coefficient c_n^t is forecast to have a nominal value $\bar{c}_n^t$ along with a bounded uncertainty of $\tilde{c}_n^t$; that is, $c_n^t\in[\bar{c}_n^t,\bar{c}_n^t+\tilde{c}_n^t]$. A naive approach would just replace c_n^t in (13) with $\bar{c}_n^t+\tilde{c}_n^t$. However, this may be overly conservative, as it will be rare that all uncertain parameters simultaneously take the extreme values.

The approach in [16] is to tune the trade-off between robustness and conservatism by confining the uncertainty to

$$\mathcal{U}_C := \left\{ \{c_n^t\in[\bar{c}_n^t,\bar{c}_n^t+\tilde{c}_n^t]\}_{n,t} \,\middle|\, \sum_{n,t}\frac{c_n^t-\bar{c}_n^t}{\tilde{c}_n^t}\le\Gamma_C \right\} \tag{17}$$

where Γ_C is the tuning parameter. Setting $\Gamma_C=0$ recovers the nominal formulation, while $\Gamma_C\ge\sum_t N_C^t$ corresponds to the naive worse-case approach. Under $\mathcal{U}_C$, the robust version is given by

$$\min \sum_{t\in\mathcal{T}}\sum_{n=1}^{N_C^t}\left[\bar{c}_n^t x_n^t + q_n^t\right] + z_C\Gamma_C - \sum_{a\in\mathcal{A}} U_a(\mathbf{p}_a) \tag{18a}$$

$$\text{over } \mathbf{p}_a\in\mathcal{P}_a, g^t\in\mathcal{G}, r^t\in\mathcal{R}, l^t, b^t\in\mathcal{B},\ t\in\mathcal{T} \tag{18b}$$

$$x_n^t\ge 0, q_n^t\ge 0,\ n=1,\ldots,N_C^t,\ t\in\mathcal{T},\ z_C\ge 0$$

$$\text{subj. to (12b)–(12d) and (14)–(16)} \tag{18c}$$

$$q_n^t + z_C \ge \tilde{c}_n^t x_n^t,\ n=1,\ldots,N_C^t,\ t\in\mathcal{T}. \tag{18d}$$

B. Robustness to Renewable Generation

The amount of renewable energy w^t naturally contains uncertainty, which can be modeled as w^t taking values in an interval $[\bar{w}^t - \tilde{w}^t, \bar{w}^t]$. Similar to (17), the desired uncertainty region for $\{w^t\}$ is

$$\mathcal{U}_W := \left\{ \{w^t \in [\bar{w}^t - \tilde{w}^t, \bar{w}^t]\}_{t\in\mathcal{T}} \left| \sum_{t\in\mathcal{T}} \frac{\bar{w}^t - w^t}{\tilde{w}^t} \le \Gamma_W \right. \right\}. \tag{19}$$

The issue is that the technique in [16] effects the protection *constraint-wise*. Since (12c) involves one uncertain parameter per constraint, the robust counterpart under $\mathcal{U}_W$ with $\Gamma_W \ge 1$ simply amounts to replacing w^t in (12c) to $\bar{w}^t - \tilde{w}^t$, which yields the worst-case solution. In principle, what is desired is the min-max approach: first maximize w.r.t. $\{w^t\} \in \mathcal{U}_W$, and then minimize the resulting objective w.r.t. the rest of the optimization variables in (12). Unfortunately, such yields a difficult nonconvex problem.

A number of alternatives for bypassing this hurdle exist. One could consider cumulative sums of the constraints in (12c)

$$\sum_{\bar{t}=1}^{t} \left[r^{\bar{t}} + l^{\bar{t}} \right] \le \sum_{\bar{t}=1}^{t} \left[g^{\bar{t}} + w^{\bar{t}} \right], \quad t \in \mathcal{T} \tag{20}$$

and apply the robust formulation with the tuning parameters Γ_W^t that gradually increase with t [21]. A more systematic (albeit more complex) approach based on Benders decomposition is proposed in [18].

Here, a simple heuristic is adapted from [19]. The idea is to first consider a single sum of all the constraints in (12c) across entire $\mathcal{T}$, and formulate the robust counterpart:

$$\sum_{t\in\mathcal{T}} \left[r^t + l^t \right] \le \sum_{t\in\mathcal{T}} \left[g^t + \bar{w}^t - s^t \right] - z_W \Gamma_W \tag{21}$$

$$s^t + z_W \ge \tilde{w}^t,\ t \in \mathcal{T} \tag{22}$$

$$s^t \ge 0,\ t \in \mathcal{T},\ z_W \ge 0. \tag{23}$$

Since the sum constraint does not guarantee the individual constraints for $t \in \mathcal{T}$ to be satisfied, an additional set of constraints are imposed as

$$r^t + l^t \le g^t + \bar{w}^t - s^t - z_W \frac{\Gamma_W}{T}, \quad t \in \mathcal{T} \tag{24}$$

which is intuitively a "disintegrated" version of the sum constraint (21). The constraints give some protection to the individual constraints (12c) via $\{s^t\}$ and z_W associated with (21), where the total protection is controlled through Γ_W.

The overall formulation for robust DR is thus given by

$$\min \sum_{t\in\mathcal{T}} \sum_{n=1}^{N_C^t} \left[\bar{c}_n^t x_n^t + q_n^t \right] + z_C \Gamma_C - \sum_{a\in\mathcal{A}} U_a(\mathbf{p}_a) \tag{25a}$$

$$\text{over } \mathbf{p}_a \in \mathcal{P}_a, g^t \in \mathcal{G}, r^t \in \mathcal{R}, l^t, b^t \in \mathcal{B}, s^t \ge 0,\ t \in \mathcal{T}$$
$$x_n^t \ge 0, q_n^t \ge 0,\ n = 1, \ldots, N_C^t,\ t \in \mathcal{T}$$
$$z_C \ge 0, z_W \ge 0 \tag{25b}$$

$$\text{subj. to (12b), (12d), (14)–(16), (18d), (21)–(22), and (24)}$$
$$0 \le r^t + l^t, t \in \mathcal{T}. \tag{25c}$$

Problems (12) and (25) are mixed-integer programs that can be solved in principle using dynamic programming (DP). However, as they are coupled across different devices through (12b), they cannot be solved separately per individual device. Thus, complexity grows rapidly as the number of devices increases, posing a major scalability hurdle. In what follows, a Lagrange relaxation approach is adopted, which enables parallelized distributed solution of the DR problems.

IV. Scalable Distributed Solution

The formulated optimization problems possess separable structures in the sense that the objectives in (12a) and (25a) as well as the coupling constraints (12b) can be represented as the sums of separate terms, where each term involves variables associated only with the corresponding device. When an optimization problem exhibits a separable structure, the Lagrange relaxation method can be adopted to decompose the problem into smaller subproblems that can be solved independently, coordinated by the Lagrange multipliers [22, Ch. 6]. Moreover, the duality gap tends to diminish as the number of separable terms increases, allowing the dual method to obtain a near-optimal solution.[1] Such an approach has been applied successfully in various domains including unit commitment in power systems [23], [27], and resource allocation in communication networks [25].

A. Lagrange Relaxation Approach

Consider first the non-robust formulation (12). Introducing multipliers $\boldsymbol{\lambda} := [\lambda^1, \ldots, \lambda^T]$, the partial Lagrangian with the coupling constraint (12b) relaxed can be written as

$$\mathcal{L}(\mathbf{p}_a, \{b^t\}, \{g^t\}, \{r^t\}, \{l^t\}; \boldsymbol{\lambda})$$
$$= \sum_{t\in\mathcal{T}} C^t(g^t) - \sum_{a\in\mathcal{A}} U_a(\mathbf{p}_a) - \sum_{t\in\mathcal{T}} \lambda^t \left(l^t - \sum_{a\in\mathcal{A}} p_a^t \right) \tag{26}$$
$$= \sum_{t\in\mathcal{T}} \left[C^t(g^t) - \lambda^t l^t \right] + \sum_{a\in\mathcal{A}} \left[\sum_{t\in\mathcal{T}} \lambda^t p_a^t - U_a(\mathbf{p}_a) \right]. \tag{27}$$

Thus, the dual function is given by

$$\mathcal{D}(\boldsymbol{\lambda}) = \min_{\substack{\mathbf{p}_a\in\mathcal{P}_a, b^t\in\mathcal{B}, \\ g^t\in\mathcal{G}, r^t\in\mathcal{R}, l^t}} \mathcal{L}(\mathbf{p}_a, \{b^t\}, \{g^t\}, \{r^t\}, \{l^t\}; \boldsymbol{\lambda}) \tag{28}$$
$$\text{subj. to (12c)–(12d).}$$
$$= \mathcal{D}_0(\boldsymbol{\lambda}) + \sum_{a\in\mathcal{A}} \mathcal{D}_a(\boldsymbol{\lambda}), \quad \text{where} \tag{29}$$

$$\mathcal{D}_0(\boldsymbol{\lambda}) := \min_{b^t\in\mathcal{B}, g^t\in\mathcal{G}, r^t\in\mathcal{R}, l^t} \sum_{t\in\mathcal{T}} \left[C^t(g^t) - \lambda^t l^t \right] \tag{30}$$
$$\text{subj. to (12c)–(12d)}$$

$$\mathcal{D}_a(\boldsymbol{\lambda}) := \min_{\mathbf{p}_a\in\mathcal{P}_a} \left[\sum_{t\in\mathcal{T}} \lambda^t p_a^t - U_a(\mathbf{p}_a) \right],\ a \in \mathcal{A}. \tag{31}$$

[1] The exact claims and proofs can be found in various forms [23]–[25], but the most comprehensive version seems to be the one in [26, Sec.5.6.1]. In essence, the claim is that under certain conditions, the duality gap of an optimization problem involving separable objectives and constraints, is proportional to the number of constraints in the primal problem, but not dependent on the number of the separable terms; see also Remark 2. Thus, as the number of terms grows, the duality gap averaged over the number of terms vanishes.

Thus, given $\boldsymbol{\lambda}$, the overall problem decomposes to a subproblem related to the grid and the distributed energy resources, and a set of subproblems, each corresponding to a DR-participating device. In fact, $\boldsymbol{\lambda}$ can be interpreted as a pricing signal that coordinates energy procurement and consumption. Specifically, $\mathcal{D}_0(\boldsymbol{\lambda})$ can be viewed as minimizing the net cost of power, which is the cost paid to the grid minus total revenue $\sum_{t\in\mathcal{T}}\lambda^t l^t$ from serving energy to devices; cf. (30). Similarly, the per-device subproblems $\mathcal{D}_a(\boldsymbol{\lambda})$ trade-off energy cost for the utility of using the devices; cf. (31).

Subproblem (30) is convex, and in fact, a linear program when $C^t(\cdot)$ is piecewise linear convex, for which efficient solvers are available. Subproblems (31) are mixed-integer programs. However, as each subproblem concerns only one device, it is much more tractable than the original coupled version. When $U_a(\mathbf{p}_a)$ can be expressed as a sum of per-time utilities, as was assumed in (9)–(10), the problems are DPs, solvable by standard techniques such as backward induction. Closed form solutions were obtained for very simple task models in [12].

In the case of using the robust formulation in (25), the per-device subproblems (31) remain unchanged. The coordinator's problem in (30), on the other hand, is replaced by

$$\min_{\substack{b^t\in\mathcal{B},g^t\in\mathcal{G},r^t\in\mathcal{R},l^t,s^t\geq 0\\ x_n^t\geq 0,q_n^t\geq 0,z_C\geq 0,z_W\geq 0}} \sum_{t\in\mathcal{T}}\left[\sum_{n=1}^{N_C^t}\left(\bar{c}_n^t x_n^t+q_n^t\right)-\lambda^t l^t\right]+z_C\Lambda_C \tag{32}$$

s. to: (12b), (12d), (14)–(16), (18d), (21)–(22), (24) and (25c).

B. Dual Problem

To obtain the optimal Lagrange multiplier $\boldsymbol{\lambda}$, the dual problem needs to be solved, where $\mathcal{D}(\boldsymbol{\lambda})$ is maximized. The goal is to recover primal solutions from the optimal dual variables. However, due to the nonzero duality gap, the primal solution recovered as an optimizer of (30)–(31) at the dual optimal solution does not necessarily satisfy the constraints of the primal problem.

In [23], the exponential method of multipliers was employed to solve not only the dual of a unit commitment problem, but also the "dual-to-dual" as a byproduct. That is, the preferred method is the one that provides not only the optimal dual variables, but also the solution to the convexified primal problem, which can help obtain a primal feasible solution through an easily implementable heuristic rule. Since the duality gap diminishes as the number of devices grows, the so-obtained primal feasible solution is close to the optimal solution when a large-scale problem is concerned. However, the method in [23] is applicable only to piecewise concave (including linear) objectives and constraints, and requires enumeration of all extreme points. This may not be applicable or straightforward for the problem at hand.

An alternative approach is to use the proximal bundle method [24], [28]. The method can handle non-differentiable objectives and yields solutions to the convexified problem at no additional cost. Consider a separable problem of the form

$$\min_{\{\mathbf{z}_a\in\mathcal{Z}_a\}}\sum_{a=0}^{A} g_a(\mathbf{z}_a) \tag{33a}$$

$$\text{subject to}\ \sum_{a=0}^{A} h_a^t(\mathbf{z}_a)\leq 0,\ t=1,\ldots,T. \tag{33b}$$

Then, based on the Lagrangian dual, the method in [24] yields the solution to the following "convexified" problem[2]

$$\min_{\{\nu_a^j\geq 0\},\{\mathbf{z}_a^j\in\mathcal{Z}_a\}}\sum_{a=0}^{A}\sum_{j=1}^{T+1}\nu_a^j g_a(\mathbf{z}_a^j) \tag{34a}$$

$$\text{subject to}\ \sum_{a=0}^{A}\sum_{j=1}^{T+1}\nu_a^j h_a^t(\mathbf{z}_a^j)\leq 0,\ t=1,\ldots,T \tag{34b}$$

$$\sum_{j=1}^{T+1}\nu_a^j=1,\ \forall a. \tag{34c}$$

If one views the variable ν_a^j as the probability of operating unit a using "mode" $\mathbf{z}_a^j$, the solution to (34a)–(34c) is seen to minimize the expectation of the original objective, while satisfying the constraints in the average sense. Moreover, it turns out that the number of units with fractional ν_a^j is at most T [24], [26], and oftentimes smaller in practice. Therefore, for large-scale problems in which $A\gg T$, the solutions for only a few units need to be perturbed to obtain a near-optimal feasible solution to the original problem (33).

The correspondence between the model problem (33) and the DR problems at hand can be easily identified by inspection. For example, for problem (12), consider defining first $\mathcal{L}:=[0,L_{\max}]$ where $L_{\max}:=\sum_{a\in\mathcal{A}}P_{\max,a}$. Then, $\mathbf{z}_a$ and $\mathcal{Z}_a$ for $a=0,1,\ldots,A$ can be identified as

$$\begin{aligned}\mathbf{z}_0 &\leftrightarrow [\{l^t\}_{t=1}^T,\{b^t\}_{t=1}^T,\{g^t\}_{t=1}^T,\{r^t\}_{t=1}^T]\\ &\in \mathcal{Z}_0\leftrightarrow\mathcal{L}^T\times\mathcal{B}^T\times\mathcal{G}^T\times\mathcal{R}^T\end{aligned} \tag{35}$$

$$\mathbf{z}_a\leftrightarrow[\{p_a^t\}_{t=1}^T]\in\mathcal{Z}_a\leftrightarrow(\mathcal{P}_a)^T,\ a=1,2,\ldots,A. \tag{36}$$

Moreover, $g_0(\cdot)\leftrightarrow\sum_{t\in\mathcal{T}}C^t(g^t)$, $h_0^t(\cdot)\leftrightarrow -l^t$, $g_a(\cdot)\leftrightarrow -U_a(\mathbf{p}_a)$ and $h_a^t(\cdot)\leftrightarrow p_a^t$ for $a\in\mathcal{A}$.

Remark 2: A set of sufficient conditions for the vanishing duality gap claim can be easily verified for (12). Three conditions are given in [26, Sec. 5.6.1,]: c1) The problem must be feasible; c2) for each $a=0,1,\ldots,A$, the set $\{(\mathbf{z}_a,\{h_a^t(\mathbf{z}_a)\}_{t=1}^T,\{g_a(\mathbf{z}_a)\}_{t=1}^T)|\mathbf{z}_a\in\mathcal{Z}_a\}$ must be compact; and c3) for each $a=0,1,\ldots,A$, for any $\tilde{\mathbf{z}}\in\mathrm{co}\mathcal{Z}_a$, where co

[2]Problem (34) is not convex per se, but its optimal objective can be shown to be the same as that of the convex problem $\inf\{\zeta_0|(\{\mathbf{z}_a\},\zeta_0,\mathbf{0})\in\overline{\mathrm{co}}\,\mathcal{V}\}$ where $\mathcal{V}:=\{(\{\mathbf{z}_a\},\zeta_0,\{\zeta^t\})\big|\{\mathbf{z}_a\in\mathcal{Z}_a\},\zeta_0\geq\sum_{a=0}^A g_a(\mathbf{z}_a),\zeta^t\geq\sum_{a=0}^A h_a^t(\mathbf{z}_a),t\in\mathcal{T}\}$ and $\overline{\mathrm{co}}\,\mathcal{V}$ denotes the closed convex hull of $\mathcal{V}$[29].

denotes the convex hull operator, there exists $\mathbf{z} \in \mathcal{Z}_a$ such that $h_a^t(\mathbf{z}) \leq \tilde{h}_a^t(\tilde{\mathbf{z}})$, where $\mathcal{Z}_a \subset \mathbb{R}^{n_a}$, and

$$\tilde{h}_a^t(\tilde{\mathbf{z}}) := \inf \left\{ \sum_{j=1}^{n_a+1} \nu^j h_a^t(\mathbf{z}^j) \middle| \tilde{\mathbf{z}} = \sum_{j=1}^{n_a+1} \nu^j \mathbf{z}^j, \mathbf{z}^j \in \mathcal{Z}_a, \sum_{j=1}^{n_a+1} \nu^j = 1,\ \nu^j \geq 0 \right\}. \quad (37)$$

Condition c1) holds trivially. Condition c2) is met if $C^t(\cdot)$ and $U_a(\cdot)$ are continuous. As for c3), the condition is again readily satisfied by considering $l^t = 0$ and $p_a^t = 0$ for all t.

C. Recovery of Primal Feasible Solution

As noted in Section II-A, once the ON/OFF profiles $\{\mathbf{u}_a\}$ of all devices are determined, the feasible power sets $\{\mathcal{P}_a\}$ become convex, and thus optimal $\{\mathbf{p}_a\}$ can be obtained by solving *convex* program (12) (or (25)). As was discussed in Section IV-B, the proximal bundle method yields a set of $\{\mathbf{u}_a^j\}$ with the fractions $\{\nu_a^j\}$, which solve (34). Thus, a simple heuristic of coin toss is adopted, where $\mathbf{u}_a^j$ is employed with probability ν_a^j for each a. In our numerical tests, this always yielded a feasible solution. Since most of the devices obtain a single $\mathbf{u}_a^j$ with $\nu_a^j = 1$ when $A \gg T$, such a heuristic provides a solution close to the optimum, with the chance of affecting feasibility being slim.

Remark 3: The heuristic does not guarantee finding a feasible solution. In case of failure, a branch-and-bound procedure can be employed in principle, where Lagrange relaxation can again be useful [30]. A more practical approach, especially for real-time DR, may be to presume infeasibility upon failure and consider load shedding.

D. Parallelized Distributed Implementation

In the envisioned architecture, the DR aggregator communicates through the advanced metering interface (AMI) of the smart grid with the subscribing entities to solve the overall DR problem in a distributed fashion. The procedure with the associated signal exchanges can be described as follows. The DR aggregator comes up with candidate Lagrange multipliers $\hat{\boldsymbol{\lambda}}^k$ at iteration k, and solve the coordinator's subproblem (30) for $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}^k$. The multipliers are broadcast to individual devices. Each device $a \in \mathcal{A}$ solves the per-device subproblems (31) with $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}^k$, and reports the optimal power profiles $\mathbf{p}_a(\hat{\boldsymbol{\lambda}}^k)$ as well as the scalar $\mathcal{D}_a(\hat{\boldsymbol{\lambda}}^k)$. Then, the coordinator updates the Lagrange multipliers and repeats the procedure until convergence. The detailed algorithm is provided in Table I; see also [28] and [24], where it is shown that the algorithm converges to the optimal Lagrange multipliers that solve the dual problem.

Upon convergence, the DR coordinator transmits $\{\nu_a^j\}$ to each device a. Each device a can now choose the ON/OFF schedule $\mathbf{u}_a$ based on a coin toss, as discussed in Section IV-C. Then, a feasible solution is found by solving a convex problem with $\{\mathbf{u}_a\}$ fixed. Note that this can be accomplished also in a distributed fashion by means of the proximal bundle method, or other methods [2], [31].

TABLE I
OVERALL DISTRIBUTED ALGORITHM.

Parameters: $\boldsymbol{\lambda}^1 \geq \mathbf{0}$, $\kappa \in (0,1)$, $\eta^1 \in [\eta_{\min}, \eta_{\max}]$

Step 0: Set $\hat{\boldsymbol{\lambda}}^1 = \boldsymbol{\lambda}^1$, $J_a^1 = \{1\}$ for all $a \in \mathcal{A}$, and $k = 1$.
Do [Subroutine 1] with $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}^1$
and obtain $\{\boldsymbol{\gamma}_a^1, \mathcal{D}_a(\hat{\boldsymbol{\lambda}}^1)\}_{a=0}^A$, and $\mathcal{D}(\hat{\boldsymbol{\lambda}}^1)$.

Step 1: Coordinator solves:
$\mathbf{d}^*, \{\xi_a^*\} = \arg \min_{\mathbf{d} \geq -\boldsymbol{\lambda}^k, \{\xi_a\}_{a=0}^A} \frac{1}{2}\eta^k ||\mathbf{d}||^2 - \sum_{a=0}^A \xi_a$
subj. to ν_a^j: $\mathcal{D}_a(\boldsymbol{\lambda}^j) + \boldsymbol{\gamma}_a^j \cdot (\boldsymbol{\lambda}^k + \mathbf{d} - \boldsymbol{\lambda}^j) \geq \xi_a$, $j \in J_a^k$, $a \in \mathcal{A} \cup \{0\}$
where $\{\nu_a^j\}$ are the associated Lagrange multipliers.
Set $\hat{\boldsymbol{\lambda}}^{k+1} = \boldsymbol{\lambda}^k + \mathbf{d}^*$.

Step 2: If $v^k := \sum_{a=0}^A \xi_a^* - \mathcal{D}(\boldsymbol{\lambda}^k) \leq 0$, go to Step 6.

Step 3: Do [Subroutine 1] with $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}^{k+1}$
and obtain $\{\boldsymbol{\gamma}_a^{k+1}, \mathcal{D}_a(\hat{\boldsymbol{\lambda}}^{k+1})\}_{a=0}^A$, and $\mathcal{D}(\hat{\boldsymbol{\lambda}}^{k+1})$.

Step 4: If $\mathcal{D}(\hat{\boldsymbol{\lambda}}^{k+1}) \geq \mathcal{D}(\boldsymbol{\lambda}^k) + \kappa v^k$, then:
Set $\boldsymbol{\lambda}^{k+1} = \hat{\boldsymbol{\lambda}}^{k+1}$ (and save $\mathcal{D}(\boldsymbol{\lambda}^{k+1})$ and $\boldsymbol{\gamma}^{k+1}$).
Choose $J_a^{k+1} \supset \{j \in J_a^k : \nu_a^j > 0\} \cup \{k+1\}$ for all $a \in \mathcal{A} \cup \{0\}$.
Else: Set $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k$ and
choose $\eta^{k+1} \in [\eta_{\min}, \eta_{\max}]$ with $\eta^{k+1} \geq \eta^k$.

Step 5: Set $k \leftarrow k+1$ and go to Step 1.

Step 6: Coordinator transmits J_a^k and $\{\nu_a^j\}_{j \in J_a^k}$ to each device $a \in \mathcal{A}$.
Each device tosses a coin and chooses $\mathbf{u}_a^j$ with probability ν_a^j.

Step 7: Solve the convex DR problem (12) or (25) with fixed $\{\mathbf{u}_a\}$ in a distributed fashion.

[Subroutine 1] Input: $\boldsymbol{\lambda}$, Output: $\{\boldsymbol{\gamma}_a, \mathcal{D}_a(\boldsymbol{\lambda})\}_{a=0}^A$ and $\mathcal{D}(\boldsymbol{\lambda})$

Coordinator solves (30) and obtains optimal $\mathcal{D}_0(\boldsymbol{\lambda})$ and $\mathbf{l}(\boldsymbol{\lambda}) := \{l^t\}$.
Coordinator broadcasts $\boldsymbol{\lambda}$ to each device $a \in \mathcal{A}$.
Each device $a \in \mathcal{A}$ solves (31), saves optimal $\mathbf{u}_a$ as $\mathbf{u}_a^k$, and feeds back $\mathcal{D}_a(\boldsymbol{\lambda})$ and $\mathbf{p}_a(\boldsymbol{\lambda})$.
Output $\boldsymbol{\gamma}_0 := -\mathbf{l}(\boldsymbol{\lambda})$, $\boldsymbol{\gamma}_a := \mathbf{p}_a(\boldsymbol{\lambda}), a \in \mathcal{A}$, and $\mathcal{D}(\boldsymbol{\lambda}) := \sum_{a=0}^A \mathcal{D}_a(\boldsymbol{\lambda})$.

Remark 4: Thanks to the parallel and distributed architecture, the algorithm can scale when the number of DR devices grows large, since the computational burden becomes distributed across the entire system. Note that while it is true that the DR aggregator needs to solve quadratic programs (QPs) of commensurate dimensions, the run time of the QP solver increases quite modestly with the dimension in practice. In terms of time complexity, it is known that the proximal bundle method can find an ϵ-optimal solution to a convex problem with at most $O(1/\epsilon^3)$ iterations [28]. The complexity of evaluating the dual objective and the subgradients is linear in A due to the dual decomposition. The QP inside the proximal bundle method can also be solved in polynomial time [32]. Thus, the overall complexity is polynomial in A. Furthermore, it is a well-documented merit of Lagrange relaxation techniques that high-quality solutions can be attained within a small number of iterations. This is also observed in our numerical tests; see Fig. 4 in Section VI and the associated discussion.

V. REAL-TIME DR

The DR algorithm discussed so far is suitable for planning the energy consumption ahead of the time of actual electricity use. At the time of use, the forecast parameters used in the DR scheduling are realized. Thus, it may be prudent to adjust the pre-planned schedule in real time by accommodating the newly acquired information to increase economic benefit.

To show how the present optimization framework can be extended to real-time DR, consider the formalism of [7]. Given a time horizon $\mathcal{T}$, let $\tau \in \mathcal{T}$ denote the current time index. By the beginning of the τth time interval, the device demands and the amount of renewable generation w^t up to time $t = \tau$ have materialized. Also, the DR scheduler has been informed about the exact electricity cost $C^t(\cdot)$ up to current time τ. However, the cost as well as the renewable resources for $t > \tau$ are still subject to forecasting errors. The idea for the real-time extension is to adopt a receding horizon approach. That is, a robust DR problem is solved at time $t = \tau$ for the remaining DR horizon $\{\tau, \tau+1, \ldots, T\}$, but the solution is applied only for the current time slot τ, which is repeated in the next slot. To formulate a robust real-time DR problem at time τ, define $\mathbf{p}_a^\tau := [p_a^\tau, p_a^{\tau+1}, \ldots, p_a^T]$ for $a \in \mathcal{A}$. The set of admissible $\mathbf{p}_a^\tau$ can also be defined appropriately. Specifically, for $a \in \mathcal{A}_1$, $\mathcal{P}_a^\tau$ can be defined as (1)–(3), but with $t = \tau, \ldots, T$. Likewise, $\mathcal{P}_a^\tau$ for $a \in \mathcal{A}_2$ is defined by additionally imposing [cf. (4)]

$$\sum_{t=\tau}^{T} p_a^t \geq E_{\text{tot},a}^\tau \tag{38}$$

where $E_{\text{tot},a}^\tau := E_a^{\text{tot}} - E_a^{\tau-1}$ represents the energy remaining to be scheduled at the beginning of the τth interval. For $a \in \mathcal{A}_3$, one adds [cf. (5)]

$$u_a^t = 1 \text{ if } \sum_{\bar{t}=\tau}^{t-1} p_a^t < E_{\text{tot},a}^\tau \text{ and } u_a^{t-1} = 1, \quad t = \tau, \ldots, T.$$

Let Γ_C^τ and Γ_W^τ denote the counterparts of Γ_C and Γ_W, respectively, for robust DR at time τ. The relevant optimization problem for real-time DR at time τ can now be written as

$$\min \sum_{n=1}^{N_C^\tau} c_n^\tau x_n^\tau + \sum_{t=\tau+1}^{T} \sum_{n=1}^{N_C^t} \left[\bar{c}_n^t x_n^t + q_n^t\right] + z_C \Gamma_C^\tau - \sum_{a\in\mathcal{A}} \sum_{t=\tau}^{T} U_a^t(p_a^t) \tag{39a}$$

$$\text{s. to (12b) and (12d) [with } t = \tau, \ldots, T] \tag{39b}$$

$$(14)\text{–}(16)\ [t = \tau, \ldots, T] \tag{39c}$$

$$(18\text{d})\ [t = \tau+1, \ldots, T] \tag{39d}$$

$$0 \leq r^\tau + l^\tau \leq g^\tau + w^\tau \tag{39e}$$

$$\sum_{t=\tau+1}^{T} [r^t + l^t] \leq \sum_{t=\tau+1}^{T} [g^t + \bar{w}^t - s^t] - z_W \Gamma_W^\tau \tag{39f}$$

$$s^t + z_W \geq \tilde{w}^t,\ t = \tau+1, \ldots, T \tag{39g}$$

$$0 \leq r^t + l^t \leq g^t + \bar{w}^t - s^t - \frac{z_W \Gamma_W^\tau}{T-\tau},\quad t = \tau+1, \ldots, T \tag{39h}$$

$$\text{over } \mathbf{p}_a^\tau \in \mathcal{P}_a^\tau$$

$$g^t \in \mathcal{G}, r^t \in \mathcal{R}, l^t, b^t \in \mathcal{B},\ t = \tau, \ldots, T$$

$$x_n^t \geq 0,\ n = 1, \ldots, N_C^t - 1,\ t = \tau, \ldots, T$$

$$q_n^t \geq 0,\ n = 1, \ldots, N_C^t, s^t \geq 0,\quad t = \tau+1, \ldots, T$$

$$z_C \geq 0, z_W \geq 0. \tag{39i}$$

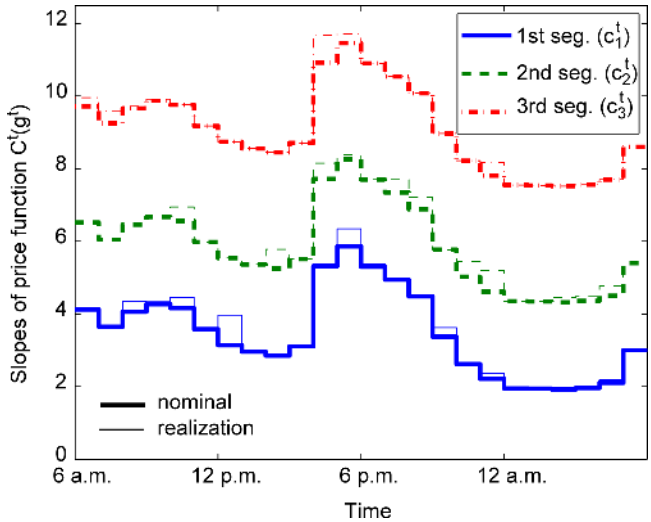

Fig. 1. Electricity price.

Note that this problem can be solved in a distributed fashion using the same technique as for the pre-planning problem.

The real-time DR is operated by employing the solution of only the current time slot. That is, although (39) is solved at time τ for the entire time horizon $t \geq \tau$, only the solution for the current time, e.g., $\{p_a^\tau\}_{a\in\mathcal{A}}$, is actually used. At the next time slot, (39) is solved again for the receded horizon, and the solution for time $\tau + 1$ is deployed. This is repeated until the end of the horizon at $t = T$.

VI. Numerical Experiments

The proposed DR algorithms were tested using numerical experiments. A total of 50 devices were scheduled over a $T = 24$ hour period, starting at 6 a.m. The first 15 devices are inelastic loads, for which utility $u_a^t(p)$ for all t was modeled to be piecewise concave with two pieces and $u_a^t(0) = 0$, where the first segment has a slope of 5 for $p \in [0, 0.7]$, and the second one has a slope of 4 for $p > 0.7$. The rest are elastic devices, and $D_a^t(E) = 0.01E$ was used for all t. The detailed operational conditions of the devices are given in Table II. An energy storage system with capacity $B_{\max} = 12.5$ and $R_{\min} = R_{\max} = 5$ was assumed with $G_{\max} = 50$. The electricity price $C^t(g^t)$ was modeled as a three-piece piecewise linear convex function with breakpoints at $\theta_1^t = 4.25$ and $\theta_2^t = 12.5$ for all t. The slopes $\{c_n^t\}$ of the three pieces at different times of use are depicted as the thick curves in Fig. 1, where the first segment was taken from the day-ahead price announced by Ameren Illinois Co. on Dec. 15, 2009 for a residential zone [33].

Fig. 2 shows the result of solving (12). The variation of the renewable resources is shown as the dotted curve. It can be seen that the energy is drawn from the grid only when the price is low. Also, it is observed that the high load in the evening hours is partly supported by using the energy stored during the off-peak hours. Fig. 3 depicts the power consumption profiles of the individual devices. The dual objective value at the optimum turns out to be -168.2, and the primal objective -167.6. The small gap verifies that the obtained solution is indeed very close to the optimum. Fig. 4 shows the evolution of the primal and the dual objectives, as well as the T Lagrange multipliers, as

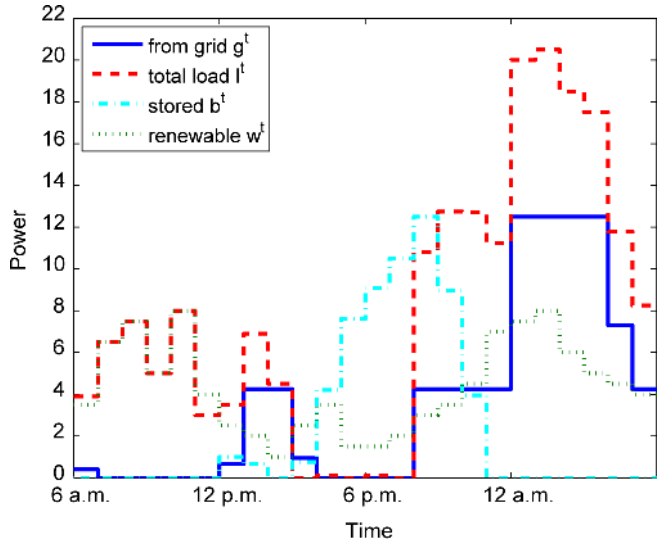

Fig. 2. Optimized power profile.

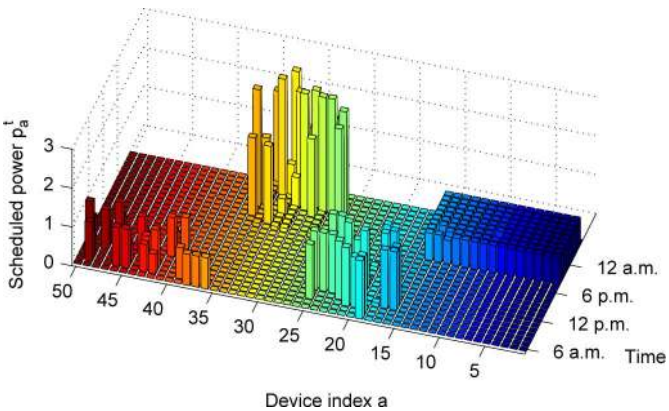

Fig. 3. Optimized power consumption of devices.

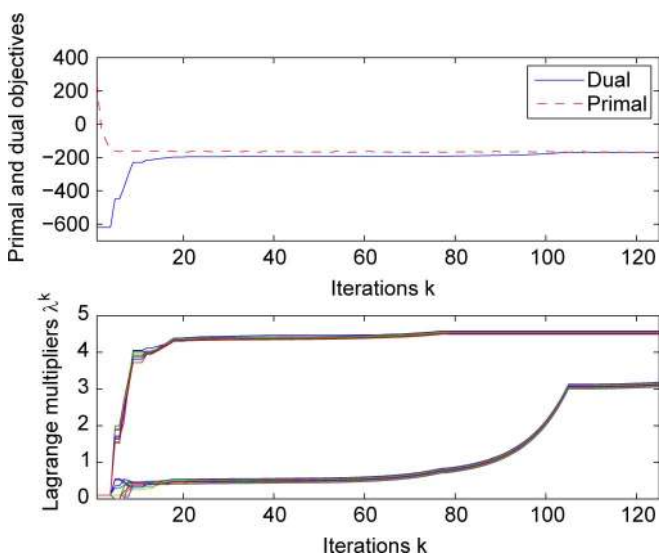

Fig. 4. Evolution of objectives and dual variables.

the iteration count grows. It can be seen that high-quality primal solutions are reached in just several iterations.

TABLE II
DEVICE CHARACTERISTICS.

class	no. of units	oper. time	$P_{\min,a}$	$P_{\max,a}$	E_a^{tot}
$\mathcal{A}_1$	15	8 p.m.-5 a.m.	0.5	1	n.a.
$\mathcal{A}_2$	10	any time	0.1	1.5	3
$\mathcal{A}_2$	10	8 p.m.-6 a.m.	0.1	3	5
$\mathcal{A}_2$	5	any time	0.8	0.8	0.8
$\mathcal{A}_3$	10	any time	0.5	1	1.5

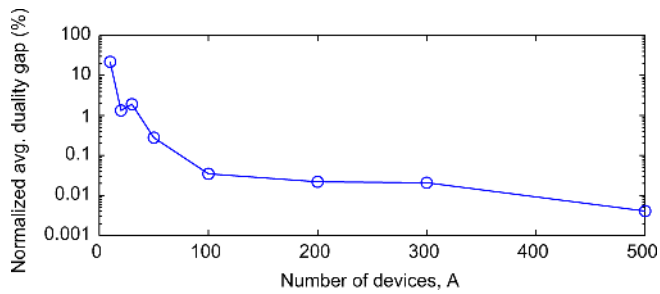

Fig. 5. Evolution of duality gaps in the number of devices.

To see how the duality gap evolves as more DR devices participate, tests were done with A varied. The proportion of devices with different characteristics was maintained as in Table II. For example, when $A = 10$, all the numbers in the second column in Table II were divided by 5. Also, the values of $G_{\max}$, $B_{\max}$, $R_{\min}$, $R_{\max}$, $\{\theta_n^t\}$ and w^t were adjusted in proportion to A. The duality gaps obtained from solving (12), averaged over 20 coin flips (cf. Section IV-C) and normalized by the dual optimal objectives are depicted in Fig. 5. It is verified that the normalized gap tends to diminish as A grows, achieving less than 0.1% of the objective when $A \geq 100$.

To verify the robust formulations, robustness to price uncertainty is first tested. Problem (25) was solved using the nominal values $\{\bar{c}_n^t\}$ as depicted (as the thick curves) in Fig. 1, and the deviations $\{\tilde{c}_n^t\}$ set to 1 for all n and t. No renewable resources or energy storage was included. The actual prices $\{c_n^t\}$ were generated by sampling from independent Gaussian distributions centered at $\bar{c}_n^t$ with standard deviation $\tilde{c}_n^t/3 = 1/3$, and then clipping the samples so that only positive deviations were allowed. One such realization is depicted as the thin curves in Fig. 1. Fig. 6 shows the histograms of the DR objectives $\sum_{t\in\mathcal{T}} C^t(g^t) - \sum_{a\in\mathcal{A}} U_a(\mathbf{p}_a)$, obtained from 100 trials. The top panel in Fig. 6 corresponds to the performance of the non-robust schedule from (12), and the bottom panel the robust version from (25) with $\Gamma_C = 1.4$. It can be seen that the robust DR achieves better (smaller) objectives, although the optimal objective 321 of the robust DR formulation is actually higher than 313 of the non-robust counterpart.

Robustness to the uncertainty of renewable sources were tested similarly. The nominal values $\{\bar{w}^t\}$ were set to what is shown in Fig. 2, while $\tilde{w}^t$ was set equal to $\bar{w}^t$ for all t. The realizations $\{w^t\}$ were generated from Gaussian distributions, followed by clipping so as to allow only negative deviations. The prices were not randomized. Since there is discrepancy between the actual and the anticipated renewable resource amounts, one has to adjust the energy drawn from the grid. We adopt a simple policy, where at each time t, g^t (and r^t if there is surplus) is adjusted to meet all scheduled loads as well as the

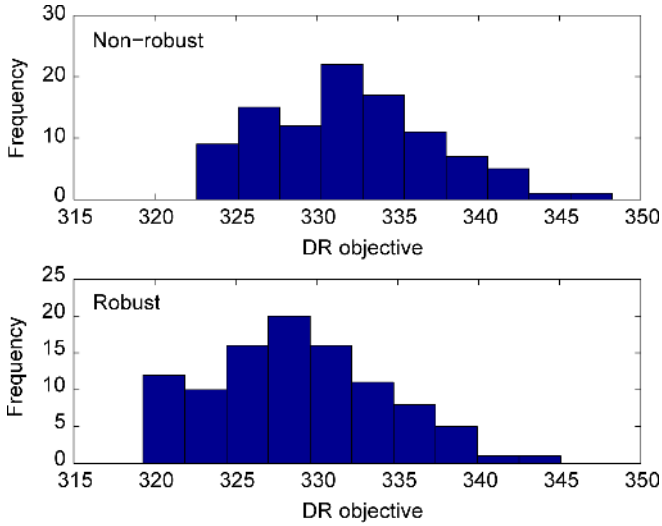

Fig. 6. Histograms of DR objectives.

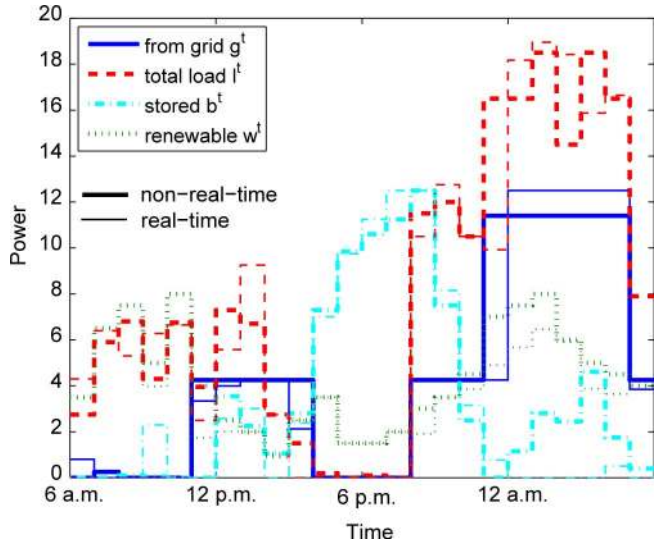

Fig. 7. Power profiles of robust and real-time DR.

planned level of stored energy. Table III presents the average DR objectives obtained from 100 trials for different values of Γ_W. Setting $\Gamma_W = 0$ corresponds to the non-robust case. It seems that Γ_W of around 10% provides the best protection. In practice, it is important to tune parameters Γ_C and Γ_W using actual data and statistics for best performance. More test results for tuning Γ_C can be found in [7]. To test the real-time DR case, a robust DR schedule with the uncertainty margins for the prices and the renewables as before, was obtained for comparison. The values $\Gamma_C = 1.4$ and $\Gamma_W = 2.4$ were used. The resultant power profiles are depicted in Fig. 7 as thick lines. Real-time adjustments were made by sequentially solving (39) with $(T-\tau)\Gamma_C/T$ and $(T-\tau)\Gamma_W/T$. The realizations of the prices and the renewable resources are shown as the thin curves in Fig. 1, and the thin dotted curves in Fig. 7, respectively. It can be seen from Fig. 7 that there was a significant drop from the forecast in the renewable resources around midnight, which the real-time DR coped with by deferring a portion of the loads. The DR objective achieved by the real-time DR was -85.3, much smaller than -78.1, achieved by trying to enforce the non-real-time schedule.

TABLE III
AVERAGE DR OBJECTIVES UNDER RENEWABLES UNCERTAINTY.

Γ_W/T	0	0.05	0.1	0.15	0.2	0.25
Avg. obj.	-90.2	-91.6	-92.8	-87.6	-78.2	-77.8

VII. CONCLUSIONS

DR formulations were considered that can optimize power consumption schedules of participating devices/subscribers, whose operating requirements may necessitate nonconvex mixed-integer models. Uncertainties in electricity prices and renewable energy resources were tackled by incorporating robustness, which can be adjusted to avoid excessive conservatism. A Lagrange relaxation approach with the proximal bundle method allowed a parallel and distributed implementation, which is advantageous for scalability and privacy. Real-time DR could be effected through a receding horizon approach. The efficacy of the proposed algorithms were verified by numerical examples.

There are other approaches for developing decentralized solutions to large-scale mixed-integer problems, such as the Dantzig-Wolfe decomposition and Benders decomposition [34]. Given our problem structure, which involves a small number of coupled constraints, employing Dantzig-Wolfe decomposition offers a viable direction. However, a standard application will require subscribers to fully disclose their utility functions to the central coordinator. Deriving custom algorithms that can mitigate such an issue, along with detailed analysis and comparison is left for future work. Performing comprehensive tests of the proposed algorithms based on real market and renewable generation data is also a topic for further study.

REFERENCES

[1] K. Hamilton and N. Gulhar, "Taking demand response to the next level," *IEEE Power Energy Mag.*, vol. 8, no. 3, pp. 61–65, May/Jun. 2010.
[2] L. Chen, N. Li, L. Jiang, and S. H. Low, "Optimal demand response: Problem formulation and deterministic case," in *Control and Optimization Methods for Electric Smart Grids*, A. Chakrabortty and M. D. Ilic, Eds. New York, NY, USA: Springer, 2012, pp. 63–85.
[3] Z. Zhu, J. Tang, S. Lambotharan, W. H. Chin, and Z. Fan, "An integer linear programming based optimization for home demand-side management in smart grid," in *Proc. IEEE PES Conf. Innovative Smart Grid Technologies (ISGT)*, Washington, D.C., USA, Jan. 2012, pp. 1–5.
[4] T. Logenthiran, D. Srinivasan, and T. Z. Shu, "Demand side management in smart grid using heuristic optimization," *IEEE Trans. Smart Grid*, vol. 3, no. 3, pp. 1244–1252, Sep. 2012.
[5] N. Gatsis and G. B. Giannakis, "Residential demand response with interruptible tasks: Duality and algorithms," in *Proc. IEEE Conf. Decision Contr.*, Orlando, FL, Dec. 2011, pp. 1–6.
[6] K. C. Sou, J. Weimer, H. Sandberg, and K. H. Johansson, "Scheduling smart home appliances using mixed integer linear programming," in *Proc. IEEE Conf. Decision Contr.*, Orlando, FL, USA, Dec. 2011, pp. 5144–5149.
[7] A. J. Conejo, J. M. Morales, and L. Baringo, "Real-time demand response model," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 236–242, Dec. 2010.

[8] P. Samadi, A.-H. Mohsenian-Rad, R. Schober, V. W. S. Wong, and J. Jatskevich, "Optimal real-time pricing algorithm based on utility maximization for smart grid," in *Proc. IEEE Intl. Conf. Smart Grid Comm.*, Gaithersburg, MD, USA, Oct. 2010, pp. 415–420.
[9] A.-H. Mohsenian-Rad and A. Leon-Garcia, "Optimal residential load control with price prediction in real-time electricity pricing environments," *IEEE Trans. Smart Grid*, vol. 1, no. 2, pp. 120–133, Sep. 2010.
[10] I. Koutsopoulos and L. Tassiulas, "Control and optimization meet the smart power grid: Scheduling of power demands for optimal energy management," in *Proc. 2nd Int. Conf. Energy-Efficient Computing Networking (e-Energy '11)*, New York, NY, USA, May 2011, pp. 41–50.
[11] M. A. A. Pedrasa, T. D. Spooner, and I. F. MacGill, "Coordinated scheduling of residential distributed energy resources to optimize smart home energy services," *IEEE Trans. Smart Grid*, vol. 1, no. 2, pp. 134–143, Sep. 2010.
[12] T. T. Kim and H. V. Poor, "Scheduling power consumption with price uncertainty," *IEEE Trans. Smart Grid*, vol. 2, no. 3, pp. 519–527, Sep. 2011.
[13] H. P. Williams, *Model Building in Mathematical Programming*. New York, NY, USA: Wiley, 1978.
[14] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. IEEE 51st Veh. Technol. Conf.-Spring*, Tokyo, Japan, May 2000, vol. 3, pp. 1854–1858.
[15] A. Ben-Tal and A. Nemirovski, "Robust optimization—methodology and applications," *Math. Program., Ser. B*, vol. 92, pp. 453–480, 2002.
[16] D. Bertsimas and M. Sim, "The price of robustness," *Oper. Res.*, vol. 52, no. 1, pp. 35–53, Jan.-Feb. 2004.
[17] D. Bertsimas and M. Sim, "Robust discrete optimization and network flows," *Math. Program., Ser. B*, vol. 98, pp. 49–71, 2003.
[18] D. Bienstock and N.Özbay, "Computing robust basestock levels," *Discrete Optimizat.*, vol. 5, no. 2, pp. 389–414, 2008.
[19] C. Bohle, S. Maturana, and J. Vera, "A robust optimization approach to wine grape harvesting scheduling," *Eur. J. Oper. Res.*, vol. 200, pp. 245–252, Jan. 2010.
[20] S. P. Bradley, A. C. Hax, and T. L. Magnanti, *Applied Mathematical Programming Reading*. Norwell, MA, USA: Addison Wesley, 1977.
[21] D. Bertsimas and A. Thiele, "A robust optimization approach to inventory theory," *Oper. Res.*, vol. 54, no. 1, pp. 150–168, Jan.-Feb. 2006.
[22] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena Scientific, 1999.
[23] D. P. Bertsekas, G. S. Lauer, N. R. Sandell, Jr, and T. A. Posbergh, "Optimal short-term scheduling of large-scale power systems," *IEEE Trans. Autom. Contr.*, vol. AC-28, no. 1, pp. 1–11, Jan. 1983.
[24] S. Feltenmark and K. C. Kiwiel, "Dual application of proximal bundle methods, including Lagrange relaxation of nonconvex problems," *SIAM J. Optimizat.*, vol. 10, no. 3, pp. 697–721, Feb./Mar. 2000.
[25] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: complexity and duality," *IEEE J. Sel. Topics Sig. Proc.*, vol. 2, no. 1, pp. 57–73, Feb. 2008.
[26] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Belmont, MA, USA: Athena Scientific, 1996.
[27] A. Frangioni, C. Gentile, and F. Lacalandra, "Solving unit commitment problems with general ramp constraints," *Intl. J. Elec. Power Syst.*, vol. 30, no. 5, pp. 316–326, Jun. 2008.
[28] K. C. Kiwiel, "Efficiency of proximal bundle methods," *J. Optim. Theory Appl.*, vol. 104, no. 3, pp. 589–603, Mar. 2000.
[29] C. Lemaréchal and A. Renaud, "A geometric study of duality gaps, with applications," *Math. Program., Ser. A*, vol. 90, pp. 399–427, 2001.
[30] A. M. Geoffrion, "Lagrangean relaxation for integer programming," *Math. Prog. Study*, vol. 2, pp. 82–114, 1974.
[31] N. Gatsis and G. B. Giannakis, "Residential load control: Distributed scheduling and convergence with lost AMI messages," *IEEE Trans. Smart Grid*, vol. 3, no. 2, pp. 770–786, Jun. 2012.
[32] Y. Ye and E. Tse, "An extension of karmarkar's projective algorithm for convex quadratic programming," *Math. Prog.*, vol. 44, no. 1–3, pp. 157–179, May 1989.
[33] Ameren Illinois Company, Day-ahead and historical prices, [Online]. Available: http://www2.ameren.com/RetailEnergy/real-timeprices.aspx
[34] G. B. Dantzig and M. N. Thapa, *Linear Programming*, ser. Springer Series in Operations Research. New York, NY, USA: Springer, 1997, vol. 2.

Seung-Jun Kim (SM'12) received the B.S. and M.S. degrees from Seoul National University, Seoul, Korea in 1996 and 1998, respectively, and the Ph.D. degree from the University of California, Santa Barbara, CA, USA, in 2005, all in electrical engineering.

From 2005 to 2008, he worked for NEC Laboratories America, Princeton, NJ, USA, as a research staff member. He is currently with the Digital Technology Center at the University of Minnesota, Minneapolis, MN, USA, where he is a Research Associate. He is also affiliated with the Department of Electrical and Computer Engineering at the University of Minnesota, where he is a Research Assistant Professor. His research interests lie in applying signal processing and optimization techniques to various domains including wireless communication and networking, smart power grids, bio and social networks.

G. B. Giannakis (F'97) received the Diploma in electrical engineering from the National Technical University of Athens, Greece, 1981, the M.Sc.degree in electrical engineering in 1983, the M.Sc. degree in mathematics in 1986, and the Ph.D. degree in electrical engineering in 1986, all from the University of Southern California (USC), Los Angeles, CA, USA.

Since 1999 he has been a professor with the Univ. of Minnesota, where he now holds an ADC Chair in Wireless Telecommunications in the ECE Department, and serves as director of the Digital Technology Center. His general interests span the areas of communications, networking and statistical signal processing—subjects on which he has published more than 350 journal papers, 580 conference papers, 20 book chapters, two edited books and two research monographs (h-index 103). Current research focuses on compressive sensing, cognitive radios, cross-layer designs, wireless sensors, social and power grid networks. He is the (co-) inventor of 21 patents issued.

Dr. Giannakis is the (co-) recipient of 8 best paper awards from the IEEE Signal Processing (SP) and Communications Societies, including the G. Marconi Prize Paper Award in Wireless Communications. He also received Technical Achievement Awards from the SP Society (2000), from EURASIP (2005), a Young Faculty Teaching Award, and the G. W. Taylor Award for Distinguished Research from the University of Minnesota. He is a Fellow of EURASIP, and has served the IEEE in a number of posts, including that of a Distinguished Lecturer for the IEEE-SP Society.