Online Learning and Distributed Control for Residential Demand Response

Xin Chen⬤ , *Student Member, IEEE*, Yingying Li⬤ , Jun Shimada⬤ , and Na Li⬤ , *Member, IEEE*

Abstract—This paper studies the automated control method for regulating air conditioner (AC) loads in incentive-based residential demand response (DR). The critical challenge is that the customer responses to load adjustment are uncertain and unknown in practice. In this paper, we formulate the AC control problem in a DR event as a multi-period stochastic optimization that integrates the indoor thermal dynamics and customer opt-out status transition. Specifically, machine learning techniques including Gaussian process and logistic regression are employed to learn the unknown thermal dynamics model and customer opt-out behavior model, respectively. We consider two typical DR objectives for AC load control: 1) minimizing the total demand, 2) closely tracking a regulated power trajectory. Based on the Thompson sampling framework, we propose an online DR control algorithm to learn customer behaviors and make real-time AC control schemes. This algorithm considers the influence of various environmental factors on customer behaviors and is implemented in a distributed fashion to preserve the privacy of customers. Numerical simulations demonstrate the control optimality and learning efficiency of the proposed algorithm.

Index Terms—Online learning, uncertain customer behavior, distributed algorithm, incentive-based demand response.

I. INTRODUCTION

DUE TO increasing renewable generation and growing peak load, electric power systems are inclined to confront a deficiency of reserve capacity. As a typical example, in mid-August 2019, Texas grid experienced record electricity demand and severe reserve emergency that were caused by the heat wave and reduced wind generation. The electricity price once soared to 9$/kWh and the Electric Reliability Council of Texas (ERCOT) issued Level 1 Energy Emergency Alert to call upon voluntary energy conservation and all available generation sources [1]. To cope with such problems, demand response (DR) is an economical and sustainable solution that strategically motivates load adjustment from end users to meet the needs of power supply [2]. In particular,

residential loads account for a large share of the total electricity usage (e.g., about 38% in the U.S. [3]), which can release significant power flexibility to facilitate system operation through a coordinated dispatch. Moreover, the widespread deployment of advanced meters, smart plugs, and built-in controllers enables the remote monitoring and control of electric appliances with two-way communications between households and load service entities (LSEs). This makes it technically feasible to implement residential DR, and well-designed DR control algorithms are necessitated to fully exploit potential flexibility.

The mechanisms for residential DR are mainly categorized as "price-based" and "incentive-based" [4]. The price-based DR programs [5] use various pricing schemes, such as time-of-use pricing, critical peak pricing, and real-time pricing, to influence and guide the electricity usage. In incentive-based DR programs [6], [7], the LSEs recruit customers to adjust their load demands in DR events with financial incentives, e.g., cash, coupon, raffle, rebate, etc. A typical residential DR event consists of two periods: 1) the *preparation period* when the LSEs anticipate an upcoming load peak or system emergency and call upon customers to participate with incentives (day-ahead or hours-ahead); 2) the *load adjustment period* when the electric appliances of participating customers are controlled to achieve certain DR goals. Meanwhile, customers are allowed to opt out (e.g., by clicking the "opt out" button in the smart phone app) if unsatisfied and override the control commands from the LSEs. In the end, the LSEs pay customers according to their actual contributions. In practice, the load adjustment period usually lasts for a few hours, and the control cycle of electric appliances varies from 5 to 30 minutes depending on the actual DR setting, which is enabled by the advanced meters with second or minute sampling rates. This paper focuses on the real-time control of electric appliances *during the load adjustment period* from the perspective of LSEs.

According to the investigations in [8], [9], offering the override (opt-out) option can greatly enhance the customers' acceptance of direct load control, which is even more effective than financial incentives. Hence, customers are generally authorized to have the opt-out option in modern residential DR programs. However, the customer opt-out behaviors are uncertain and unknown to LSEs in practice, which brings significant challenges to the real-time DR control. References [8]–[11] indicate that customer DR behaviors are influenced by *individual preference* and *environmental factors*. The individual preference relates to customers' intrinsic socio-demographic characteristics, e.g., income, education, age,

attitude to energy saving, etc. The environmental factors refer to real-time externalities such as electricity price, indoor temperature, offered incentive, weather conditions, etc. Without considering customers' actual willingness, a blind DR control scheme may lead to high opt-out rates and inefficient load adjustment.

To address the uncertainty issue, data-driven learning techniques can be employed to learn customer DR behaviors with historical data and through online interactions and observations. Comprehensive reviews on the application of reinforcement learning (RL) for DR are provided in [12], [13]. References [14]–[16] design home energy management systems to optimally schedule electric appliances under time-varying electricity price, where Q-learning is used to learn customer preferences and make rescheduling decisions. In [17], a real-time DR strategy is presented for optimal arrangement of home appliances based on deep RL and policy search algorithms, considering the uncertainty of the resident's behavior, electricity price, and outdoor temperature. Reference [18] applies batch RL to dispatch thermostatically controlled loads (TCLs) with the exploitation of exogenous data for day-ahead scheduling. Reference [19] proposes an incentive-based DR algorithm using RL and deep neural networks, to assist LSEs in designing the optimal incentive rates with uncertain energy prices and demands. In [20], [21], the multi-armed bandit method and its variants are adopted to select the right customers for DR participation at the preparation period to deal with unknown customer responses. However, for real-time load control in incentive-based DR, most existing works do not consider or overly simplify the uncertainty of customer behaviors, and the influence of various environmental factors is generally neglected. This causes significant mismatches between theory and practice. Hence, the development of real-time DR control algorithms that take into account customer opt-out behaviors remains largely unresolved.

Contribution: This paper studies the incentive-based residential DR programs that control air conditioner (AC) loads to optimize certain DR performances, e.g., minimizing the total AC load or closely tracking a target power trajectory. To this end, we propose a novel framework to model the real-time DR control as a multi-period stochastic optimization problem that integrates thermal dynamics and customer behavior transitions. In particular, Gaussian process (GP) [22] is adopted to build a non-parametric indoor thermal dynamical model from historical metering data, and logistic regression [23] is used to model the customer opt-out behaviors under the influence of environmental factors. Based on the Thompson sampling (TS) framework [24], we develop a distributed online DR control algorithm to learn the customer opt-out behaviors and make real-time AC power control decisions. The main merits of the proposed algorithm are summarized as follows.

1) The individual preferences of customers and time-varying environmental factors are taken into account, which improves the predictions on customer opt-out behaviors and leads to efficient AC control schemes.

2) This algorithm is implemented in a distributed manner and thus can be directly embedded in local household AC appliances, smart plugs, or smart phone apps. Moreover, the communication burdens are mitigated and the customer privacy can be preserved.

3) Inheriting the merits of TS, this algorithm has a convenient decomposition structure of learning and optimization, and strikes an effective balance between exploration and exploitation in the online learning process.

The remainder of this paper is organized as follows: Section II provides a preliminary introduction on GP and TS. Section III presents the optimal AC control models with the learning techniques. Section IV develops the distributed online learning and AC control algorithm. Numerical tests are carried out in Section V, and conclusions are drawn in Section VI.

II. PRELIMINARIES ON LEARNING TECHNIQUES

This section provides a preliminary introduction on the two key learning techniques used in this paper, i.e., Gaussian process and Thompson sampling.

A. Gaussian Process

Gaussian process is a non-parametric supervised machine learning method [22] that has been widely used to model nonlinear system dynamics [27]. A formal definition of GP over a function $f(\boldsymbol{x})$ is that any finite number of function realizations $(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), f(\boldsymbol{x}_3), \ldots)$ are random variables and follow a joint Gaussian distribution, which is fully specified by the mean function $m(\boldsymbol{x})$ and the (kernel) covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$. Consider learning an unknown function $y = f(\boldsymbol{x})$ based on a training dataset $\mathcal{D}$ of n noisy observations, i.e., $\mathcal{D} := \{(\boldsymbol{x}_i, \hat{y}_i) | i = 1, \ldots, n\}$. It aims to infer the function value $f(\boldsymbol{x}_*)$ for a new point $\boldsymbol{x}_*$. Denote $\hat{\boldsymbol{y}}_{\mathcal{D}} := (\hat{y}_i)_{i=1}^n$ and $\boldsymbol{f}_{\mathcal{D}} := (f(\boldsymbol{x}_i))_{i=1}^n$. By the GP definition, $(\boldsymbol{f}_{\mathcal{D}}, f(\boldsymbol{x}_*))$ are assumed to be random variables and follow a joint Gaussian distribution

$$\begin{bmatrix} \boldsymbol{f}_{\mathcal{D}} \\ f(\boldsymbol{x}_*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m}_{\mathcal{D}} \\ m(\boldsymbol{x}_*) \end{bmatrix}, \begin{bmatrix} K_{\mathcal{D},\mathcal{D}} & \boldsymbol{k}_{*,\mathcal{D}} \\ \boldsymbol{k}_{*,\mathcal{D}}^{\top} & k(\boldsymbol{x}_*, \boldsymbol{x}_*) \end{bmatrix} \right),$$

where vector $\boldsymbol{k}_{*,\mathcal{D}} := [k(\boldsymbol{x}_*, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}_*, \boldsymbol{x}_n)]^{\top}$ and $\boldsymbol{m}_{\mathcal{D}} := (m(\boldsymbol{x}_i))_{i=1}^n$. $K_{\mathcal{D},\mathcal{D}} \in \mathbb{R}^{n \times n}$ is the covariance matrix, whose ij-component is $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Conditioning on the given observations $\mathcal{D}$, it is known that the posterior distribution of $f(\boldsymbol{x}_*)|(\boldsymbol{f}_{\mathcal{D}} = \hat{\boldsymbol{y}}_{\mathcal{D}})$ is also Gaussian, i.e., $\mathcal{N}(\mu_{*|\mathcal{D}}, \sigma_{*|\mathcal{D}}^2)$, with the closed form

$$\mu_{*|\mathcal{D}} = m(\boldsymbol{x}_*) + \boldsymbol{k}_{*,\mathcal{D}}^{\top} K_{\mathcal{D},\mathcal{D}}^{-1} (\hat{\boldsymbol{y}}_{\mathcal{D}} - \boldsymbol{m}_{\mathcal{D}}), \tag{1a}$$

$$\sigma_{*|\mathcal{D}}^2 = k(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_{*,\mathcal{D}}^{\top} K_{\mathcal{D},\mathcal{D}}^{-1} \boldsymbol{k}_{*,\mathcal{D}}. \tag{1b}$$

Then the mean value $\mu_{*|\mathcal{D}}$ (1a) can be used as the prediction on $f(\boldsymbol{x}_*)$, and the variance $\sigma_{*|\mathcal{D}}^2$ (1b) provides a confidence estimate for this prediction. The merits of GP include 1) GP is a non-parametric method that avoids the bias of model selection; 2) GP works well with small datasets; 3) GP can incorporate prior domain knowledge by defining priors on hyperparameters or using a particular covariance function. The major issue of GP is the computational complexity, which scales cubically in the number of observations, i.e., $O(n^3)$.

Algorithm 1 Thompson Sampling (TS) Algorithm [24]
1: **Input:** Prior distribution $\mathcal{P}$ on θ.
2: **for** $t = 1$ to T **do**
3:    Sample $\hat{\theta} \sim \mathcal{P}$.
4:    $a_t \leftarrow \arg\max_{a \in \mathcal{A}} \mathbb{E}_{\mathcal{P}_{\hat{\theta}}}[r(z_t)\|a_t = a]$.   Apply a_t and observe z_t.
5:    Posterior update: $\mathcal{P} \leftarrow \frac{\mathcal{P}(\theta)\mathcal{P}_\theta(z_t\|a_t)}{\int_{\tilde{\theta}} \mathcal{P}(\tilde{\theta})\mathcal{P}_{\tilde{\theta}}(z_t\|a_t)\, d\tilde{\theta}}$.
6: **end for**

B. Thompson Sampling

Thompson sampling [24] is a prominent Bayesian learning framework that was originally developed to solve the multi-armed bandit (MAB) problem [25] and can be extended to tackle other online learning problems. Consider a classical T-periods MAB problem where an agent selects an action (called "arm") a_t from the action set $\mathcal{A}$ at each time $t \in \{1, 2, \ldots, T\}$. After taking action a_t, the agent observes an outcome z_t that is randomly generated from a conditional distribution $\mathcal{P}_\theta(\cdot|a_t)$, and then obtains a reward $r_t = r(z_t)$ with known reward function $r(\cdot)$. The agent is initially uncertain about the parameter θ in $\mathcal{P}_\theta$ but aims to maximize the total expected reward using the observation feedback. To achieve good performance, it is generally required to take actions with an effective balance between 1) *exploring* poorly-understood actions to gather new information that may improve future reward and 2) *exploiting* what is known for decision to maximize the immediate reward.

TS is a straightforward online learning algorithm that strikes an effective balance between exploration and exploitation. As shown in Algorithm 1, TS treats the unknown θ as a random variable and represents the initial belief on θ with a prior distribution $\mathcal{P}$. At each time t, TS draws a random sample $\hat{\theta}$ from $\mathcal{P}$ (Step 3), then takes the optimal action that maximizes the expected reward based on the sample $\hat{\theta}$ (Step 4). After outcome z_t is observed, the Bayesian rule is applied to update the distribution $\mathcal{P}$ over θ and obtain the posterior (Step 5).

The main features of the TS algorithm are listed below.

- As outcomes accumulate, the predefined prior distribution will be washed out and the posterior converges to the true distribution or concentrates on the true value of θ.
- The TS algorithm encourages exploration by the random sampling (Step 3). As the posterior distribution gradually concentrates, less exploration and more exploitation will be performed, which leads to an effective balance.
- The key advantage of the TS algorithm is that a complex online problem is decomposed into a Bayesian learning task (Step 5) and an offline optimization task (Step 4) [26], while the optimization remains the original formulation without being complicated by the learning task.

III. PROBLEM FORMULATION

Consider the residential DR program that controls AC power consumption for load adjustment, where a system aggregator (SA) interacts with N residential customers over sequential DR events. Each DR event[1] is formulated as a finite time horizon $[T] := \{1, 2, \ldots, T\}$ with the time gap Δt. Depending on the practical AC control cycle, Δt may be different (e.g., 5 minutes or 10 minutes) across DR events,[2] and the time length T could also be different (e.g., 2 hours or 3 hours). The SA aims to learn customers' opt-out behaviors and make real-time AC control decisions to optimize aggregate DR performance. In this paper, we focus on the control of household mini-split AC units, while the proposed method is applicable to heating, ventilation and air conditioning (HVAC) systems, space heaters, and other TCLs. Moreover, the proposed method works for both the AC heating and cooling cases separately.

In this section, we firstly establish the AC control model for each individual customer during a DR event, then use Gaussian process and logistic regression to learn the unknown thermal dynamics and customer opt-out behaviors, respectively. With all these in place, two system-level optimization models with different DR objectives are built for the SA to generate optimal AC power control schemes over N customers.

A. Customer-Side AC Control Model During a DR Event

1) Decision Variable: During a DR event, denote $u_{i,t}$ as the AC power consumption for customer $i \in [N] := \{1, \ldots, N\}$ at time $t \in [T]$, which is the *decision variable* and satisfies

$$0 \le u_{i,t} \le \bar{u}_i, \ \left|u_{i,t} - u_{i,t-1}\right| \le \Delta u_i, \quad \forall t \in [T] \qquad (2)$$

where $\bar{u}_i$ is the rated AC power capacity and Δu_i denotes the AC power drift limit that prevents dramatic changes. In (2), $u_{i,t}$ is continuously controllable, since it indeed denotes the average AC power during the time interval $[t-1, t]$, which can be realized by appropriately adjusting the AC cycling rate, i.e., the time ratio of the on status over Δt. Nevertheless, a discrete AC control model for customer i with

$$u_{i,t} = \delta_{i,t} \cdot \bar{u}_i, \ \delta_{i,t} \in \{0, 1\}, \quad \forall t \in [T]$$

can be applied as well, where the AC unit switches between the on ($\delta_{i,t} = 1$) and off ($\delta_{i,t} = 0$) modes.

2) Opt-Out Status and Transition: Denote binary variable $z_{i,t} \in \{0, 1\}$ as the *opt-out status* of customer i at time t, which equals 1 if customer i stays in and 0 if opts out. Initialize $z_{i,0} = 1$ for all customers at the beginning of a DR event. As mentioned above, customer opt-out behaviors are influenced by various environmental factors. Accordingly, the binary opt-out statuses $z_i := (z_{i,t})_{t \in [T]}$ of each customer $i \in [N]$ are modeled as a time series of random variables that are independent of other customers and follow the transition probability (3):

$$\mathbb{P}(z_{i,t} = 0|z_{i,t-1} = 0) = 1, \ \forall t \in [T] \qquad (3a)$$
$$\mathbb{P}(z_{i,t} = 1|z_{i,t-1} = 1) = p_i(\boldsymbol{w}_{i,t}), \ \forall t \in [T] \qquad (3b)$$

Here, $\boldsymbol{w}_{i,t}$ is the column vector that collects the environmental factors at time t, which will be elaborated in the next

---

[1]To avoid confusion, in the following text, we specifically refer to the load adjustment period when referring to "DR event."

[2]Generally, the control time gap Δt should be larger than the metering data collection period, which is enabled by current smart meters with second or minute sampling rates.
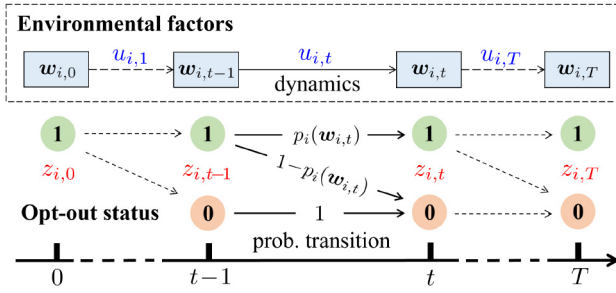
Fig. 1. Opt-out status transition and dynamics of AC control in a DR event.

part. $p_i(\boldsymbol{w}_{i,t}) \in [0, 1]$ is the transition probability function that captures how environmental factors $\boldsymbol{w}_{i,t}$ affect the customer opt-out behaviors.

As illustrated in Figure 1, equation (3a) enforces the DR rule that once customer i opts out at a certain time, this customer will remain the opt-out status for the rest of the current DR event. Equation (3b) indicates that if customer i stays in at time $t-1$, this customer may remain stay-in at next time t with probability $p_i(\boldsymbol{w}_{i,t})$, or choose to opt out with probability $1 - p_i(\boldsymbol{w}_{i,t})$. We further explain the transition model (3) with the following remark.

Remark 1: The opt-out status transition model (3) exhibits the Markov property, where the transition probability $p_i(\boldsymbol{w}_{i,t})$ is functional on the environmental factors $\boldsymbol{w}_{i,t}$ at time t. This facilitates the subsequent development and solution of the optimal AC control models, but does not sacrifice the modeling generality. Because it is free to choose suitable environmental factors, so that all the useful information is captured in $\boldsymbol{w}_{i,t}$ and the Markov property is preserved. Basically, by including all the necessary known information in the enlarged state at time t, which is known as *state augmentation*, any multi-period control problem can generally be modeled as a Markov decision process [28]. See the next part for the detailed definition and selection of the environmental factors.

3) Environmental Factors: Based on the empirical investigations in [8]–[11], we present below several key environmental factors that influence customers' opt-out behaviors. In particular, the first three factors are affected by the AC control scheme, thus their dynamics models are introduced as well.

• *(Indoor Temperature):* Denote $s_{i,t}$ as the indoor temperature for customer i at time t. The associated thermal dynamical model can be formulated as

$$s_{i,t} = f_i\big(s_{i,t-1}, s_{i,t-1}^{\text{out}}, u_{i,t}\big), \quad t \in [T] \qquad (4)$$

where $s_{i,t}^{\text{out}}$ is the outdoor temperature at time t, $(s_{i,0}, s_{i,0}^{\text{out}})$ are the initial temperatures in the beginning of the DR event, and $f_i(\cdot)$ denotes the thermal dynamics function.

• *(Accumulated Thermal Discomfort):* We define $d_{i,t}$ as the accumulated thermal discomfort for customer i at time t, and let it follow the dynamics (5) with $d_{i,0} = 0$:

$$d_{i,t} = d_{i,t-1} + \Delta t \cdot \big(\max\{s_{i,t} - s_i^{\text{set}}, 0\}\big)^2, \quad t \in [T] \qquad (5)$$

where s_i^{set} denotes the defaulted AC setting temperature by customer i. The operator $\max\{x, 0\}$ takes the larger value between x and 0, which means that only the indoor temperature

higher than the setpoint will cause thermal discomfort for the AC cooling case in the summer. Besides, the quadratic form in (5) captures that the thermal discomfort increases faster as the temperature deviation becomes larger [29].

• *(Incentive Credit):* Denote $r_{i,t}$ as the incentive credit offered to customer i at time t. We consider a general incentive scheme (6), while other incentive schemes can be used as well.

$$r_{i,t} = r_0 + r_1 \cdot \bar{u}_i \cdot t + r_2 \cdot \Delta t \sum_{\tau=1}^{t} \big|u_{i,\tau}^{\text{set}} - u_{i,\tau}\big|, \quad t \in [T]. \quad (6)$$

In (6), the first term r_0 is the base credit for DR participation. The second term is the stay-in bonus that is proportional to time t and the AC power capacity $\bar{u}_i$ with coefficient r_1. The third term is the reimbursement for the actual load adjustment with credit coefficient r_2, where $u_{i,t}^{\text{set}}$ is the associated AC power at time t to maintain the setting temperature s_i^{set}.

Essentially, $(u_{i,t}^{\text{set}})_{t \in [T]}$ can be regarded as the baseline AC power consumption of customer i when no DR control is implemented. With the thermal dynamical model (4) and given $(s_i^{\text{set}}, s_{i,t-1}^{\text{out}})$, one can compute $u_{i,t}^{\text{set}}$ by solving equation (7):

$$s_i^{\text{set}} - f_i\big(s_i^{\text{set}}, s_{i,t-1}^{\text{out}}, u_{i,t}^{\text{set}}\big) = 0, \qquad (7)$$

which follows the definition of maintaining the defaulted setting indoor temperature s_i^{set}.

Other key environmental factors that would influence customers' opt-out behaviors include real-time electricity prices, weather conditions, duration of the DR event, fatigue effect, etc. These factors are treated as given parameters that can be obtained or predicted ahead of time. Accordingly, the vector $\boldsymbol{w}_{i,t}$ in (3b) can be defined as the combination of the environmental factors mentioned above:

$$\boldsymbol{w}_{i,t} := \big(s_{i,t}, d_{i,t}, r_{i,t}, \text{electricity price at } t, \dots, \big). \qquad (8)$$

Remark 2: We note that the definition and selection of useful environment factors are complex and tricky in practice. For instance, the definition of $\boldsymbol{w}_{i,t}$ in (8) only contains the present status at time t, while the past values and future predictions may also be included in $\boldsymbol{w}_{i,t}$ to capture the temporal dependence. This is related to the feature engineering problem in machine learning, which is expected to be conducted based on real data and make a trade-off between complexity and effectiveness. Nevertheless, the proposed learning and AC control method is a general framework that is applicable to different choices of the environmental factors.

One critical issue for the residential AC control is that the thermal dynamics function $f_i(\cdot)$ in (4) and the customer opt-out behavior function $p_i(\cdot)$ in (3b) are generally unknown. To address this issue, learning techniques are used to estimate the unknown models with real data, which are presented in the following two subsections.

B. Learning for Thermal Dynamics Model

The practical implementation of AC control for residential DR is generally achieved through smart plugs or smart AC units with built-in controllers and sensors. These smart devices are able to measure, store, and communicate the temperature

and AC power data in real time. Hence, the thermal dynamics model (4) can be estimated based on fine-grained historical measurement data. To this end, we provide the following two thermal model estimation schemes.

1) Linear Model: Given a time series of historical indoor/outdoor temperature and AC power data, one can fit a classical linear thermal dynamics model (9) [30] and obtain the coefficients (κ_i, η_i) via linear regression:

$$s_{i,t} = s_{i,t-1} + \kappa_i \cdot \left(s_{i,t-1}^{\text{out}} - s_{i,t-1}\right) + \eta_i \cdot u_{i,t}, \qquad (9)$$

where coefficients κ_i and η_i specify the thermal characteristics of the room with AC and the ambient. A positive (negative) η_i indicates that AC works in the heating (cooling) mode.

2) Gaussian Process Model: An alternative scheme is to employ the Gaussian process method (introduced in Section II) to model the thermal dynamics as (10), which can capture the nonlinearity in the data pattern:

$$s_{i,t} = m\left(x_{i,t-1}\right) + k_{*,\mathcal{D}}\left(x_{i,t-1}\right)^{\top} K_{\mathcal{D},\mathcal{D}}^{-1}(\hat{y}_{\mathcal{D}} - m_{\mathcal{D}}), \quad (10)$$

where $x_{i,t-1} := (s_{i,t-1}, s_{i,t-1}^{\text{out}}, u_{i,t})$, and the notations with subscript $\mathcal{D}$ denote the corresponding terms associated with the historical (training) dataset as presented in (1a).

The main virtue of the linear model (9) lies in its simplicity of implementation and interpretability. In contrast, the non-parametric GP model (10) offers more modeling flexibility and can capture the nonlinear relation and avoid the bias of model class selection, in the cost of computational complexity. Besides, other suitable regression methods can be applied to model the thermal dynamics as well. The choice of model depends on the practical DR requirements on computational efficiency and modeling accuracy. The historical data above refer to the available datasets that have been collected by advanced meters before the DR event, thus the thermal dynamics model can be estimated in an offline fashion. Nevertheless, dynamic regression that uses the latest data to fit an updated model along the DR event is also applicable to further enhance the prediction accuracy.

C. Learning for Customer Opt-Out Behaviors

Since the customer opt-out status $z_{i,t}$ is binary, logistic regression [23] is used to model the transition probability function $p_i(w_{i,t})$ in (3b). Because the output of logistic regression is naturally a probability value within $[0, 1]$, and it is easy to implement and interpret. Moreover, logistic regression is compatible with the online Bayesian learning framework with efficient posterior update approaches (see Section IV-C for details). Accordingly, we formulate the transition probability function $p_i(w_{i,t})$ as the logistic model (11):

$$p_i\left(w_{i,t}\right) = \frac{1}{1 + \exp\left(-\left(\alpha_i + \beta_i^{\top} w_{i,t}\right)\right)}, \qquad (11)$$

where β_i is the weight vector describing how customer i reacts to the environmental factors $w_{i,t}$, and α_i depicts the individual preference. Define $\hat{w}_{i,t} := (1, w_{i,t})$ and $\theta_i := (\alpha_i, \beta_i)$. Then the linear term in (11) becomes $-\hat{w}_{i,t}^{\top}\theta_i$. Without causing any confusion, we use $p_{\theta_i}(w_{i,t})$ and $p_i(w_{i,t})$ interchangeably.

As a consequence, the unknown information of customer i's behaviors is summarized as vector θ_i, which can be estimated from the observations of $(\hat{w}_{i,t}, z_{i,t})$ in DR events. In contrast to the thermal dynamics model learning, the observation data of $(\hat{w}_{i,t}, z_{i,t})$ are not historically available but can only be obtained along with the real implementation of DR events. This leads to an *online* customer behavior learning and AC power control problem. Thus we employ the TS framework to develop the online AC control algorithm in Section IV to effectively balance exploration and exploitation.

D. System-Level Optimal AC Control Models

In a typical DR setting, once a customer opts out, the AC unit will automatically be switched back to the defaulted operation mode with the original customer-set temperature s_i^{set}. Taking the opt-out status into account, the actual AC power consumption $\hat{u}_{i,t}$ can be formulated as

$$\hat{u}_{i,t} = u_{i,t} \cdot z_{i,t-1} + u_{i,t}^{\text{set}} \cdot \left(1 - z_{i,t-1}\right), \qquad (12)$$

which equals $u_{i,t}$ if customer i stays in ($z_{i,t-1} = 1$) or $u_{i,t}^{\text{set}}$ if opts out ($z_{i,t-1} = 0$). Denote $u_i := (u_{i,t})_{t \in [T]}$, $s_i := (s_{i,t})_{t \in [T]}$, $d_i := (d_{i,t})_{t \in [T]}$, $r_i := (r_{i,t})_{t \in [T]}$, and $z := (z_i)_{i \in [N]}$.

To simplify the expression, we reformulate the AC control constraints (2), the opt-out status transition (3), (8), (11), and the dynamics of environmental factors (5), (6), (9) or (10), for customer $i \in [N]$, as the following compact form (13):

$$(u_i, z_i, s_i, d_i, r_i) \in \mathcal{X}_i, \qquad (13)$$

where $\mathcal{X}_i$ denotes the corresponding feasible set. Then two system-level optimal control (SOC) models, i.e., (14) and (15), with different DR goals are established for the SA to solve optimal AC power control schemes over N customers.

1) SOC-1 model (14) aims to reduce as much AC load as possible in a DR event, which can be used to flatten the load peaks or mitigate reserve deficiency emergency.

$$\min_{u_{i,t}} \quad \mathbb{E}_z \left[\sum_{i=1}^{N} \left[\Delta t \sum_{t=1}^{T} \hat{u}_{i,t} + \rho_i(1 - z_{i,T}) \right] \right] \qquad (14a)$$

$$\text{s.t.} \quad (u_i, z_i, s_i, d_i, r_i) \in \mathcal{X}_i, \quad \forall i \in [N] \qquad (14b)$$

where objective (14a) minimizes the expected total AC energy consumption over the DR event, plus an opt-out penalty term in the last time step. ρ_i is the penalty coefficient that can be tuned to balance load reduction and opt-out outcomes. $\mathbb{E}_z[\cdot]$ denotes the expectation that is taken over the randomness of the customer opt-out status z. Constraint (14b) collects the counterparts of (13) for all customers.

2) SOC-2 model (15) aims to closely track a regulated power trajectory $(L_t)_{t \in [T]}$, which is determined by the upper-level power dispatch or the DR market.

$$\min_{u_{i,t}} \quad \mathbb{E}_z \left[\sum_{t=1}^{T} \Delta t \left(\sum_{i=1}^{N} \hat{u}_{i,t} - L_t \right)^2 + \sum_{i=1}^{N} \rho_i\left(1 - z_{i,T}\right) \right] \quad (15a)$$

$$\text{s.t.} \quad (u_i, z_i, s_i, d_i, r_i) \in \mathcal{X}_i, \quad \forall i \in [N]. \qquad (15b)$$

Objective (15a) minimizes the expected total squared power tracking deviation from the global target L_t, plus the same

opt-out penalty term defined in (14a). Constraint (15b) is the same as (14b).

Remark 3: The penalty term $\rho_i(1 - z_{i,T})$ in the objectives (14a) and (15a) serves as the final state cost in a finite-horizon control planning problem, which is used to restrict the last control action $u_{i,T}$. Without this penalty term, the last control action $u_{i,T}$ would be too radical with no regard for the opt-out outcome at time T and lead to frequent final opt-out $z_{i,T} = 0$. Besides, this penalty term is a useful tool for the SA to make a trade-off between the DR goals and the customer opt-out results through adjusting the coefficient ρ_i.

The two SOC models (14) and (15) are indeed discrete-time finite-horizon control planning problems, which are in the form of nonconvex stochastic optimization, and the stochasticity results from the probabilistic opt-out status transition (3). We develop the distributed solution methods for the SOC models (14) and (15) in the next section.

IV. DISTRIBUTED SOLUTION AND ALGORITHM DESIGN

For the real-time AC control in a DR event, we pursue a distributed implementation manner such that
1) the control algorithm can be directly embedded in the local home electric appliances or smart phone apps;
2) heavy communication burdens between the SA and households are avoided during the DR event;
3) the private information of customers can be protected.

In this section, we propose the distributed solution methods for the SOC models (14) and (15), then develop the distributed online AC control algorithm based on the TS framework.

A. Distributed Solution of SOC-1 Model (14)

Since the opt-out status transition of one customer is assumed to be independent of other customers in (3), objective (14a) in the SOC-1 model has no substantial coupling among different customers. Hence, the SOC-1 model (14) can be equivalently decomposed into N local problems, i.e., model (16) for each customer $i \in [N]$.

$$\min_{u_{i,t}} \quad \mathbb{E}_{z_i}\left[\Delta t \sum_{t=1}^{T} \hat{u}_{i,t} + \rho_i(1 - z_{i,T})\right] \quad (16a)$$

$$\text{s.t.} \quad (\boldsymbol{u}_i, \boldsymbol{z}_i, \boldsymbol{s}_i, \boldsymbol{d}_i, \boldsymbol{r}_i) \in \mathcal{X}_i. \quad (16b)$$

The sum of objectives (16a) over all N customers is essentially objective (14a) in the SOC-1 model, and constraint (16b) is the individual version of (14b) for customer i.

The local model (16) is a stochastic optimization with the expectation over $\boldsymbol{z}_i$ in the objective. Since $\boldsymbol{z}_i$ follows the transition (3) with the probability function $p_{\boldsymbol{\theta}_i}(\boldsymbol{w}_{i,t})$ (11), we can derive the analytic form of the expectation in (16a), which leads to expression (17):

$$\sum_{t=1}^{T}\left[\Delta t(u_{i,t} - u_{i,t}^{\text{set}})\prod_{\tau=1}^{t-1} p_{\boldsymbol{\theta}_i}(\boldsymbol{w}_{i,\tau})\right] - \rho_i \prod_{t=1}^{T} p_{\boldsymbol{\theta}_i}(\boldsymbol{w}_{i,t}). \quad (17)$$

Expression (17) only differs from the expectation in (16a) by a constant term $\Delta t \sum_{t=1}^{T} u_{i,t}^{\text{set}} + \rho_i$, thus they are equivalent in optimization. See the Appendix for the detailed derivation.

When $\boldsymbol{\theta}_i$ is given, we can obtain the optimal AC control schemes $\boldsymbol{u}_i^*$ for each customer i via solving the local model (16), which is implemented in a fully distributed manner. Meanwhile, the aggregation of $\boldsymbol{u}_i^*$ over all N customers is essentially an optimal solution to the SOC-1 model (14).

B. Distributed Solution of SOC-2 Model (15)

The objective (15a) in the SOC-2 model has coupling among different customers due to tracking a global power trajectory $(L_t)_{t \in [T]}$. To solve this problem distributedly, we introduce a local tracking trajectory $\boldsymbol{l}_i := (l_{i,t})_{t \in [T]}$ for each customer $i \in [N]$ with $\sum_{i=1}^{N} l_{i,t} = L_t$ for all $t \in [T]$. Then we substitute L_t by $\sum_{i=1}^{N} l_{i,t}$ in objective (15a), so that (15a) can be approximated[3] by the decomposable form (18a), which takes the form of a sum over N customers. As a result, the SOC-2 model (15) is modified as (18):

$$\min_{u_{i,t}, l_{i,t} \geq 0} \quad \sum_{i=1}^{N} \mathbb{E}_{z_i}\left[\sum_{t=1}^{T} \Delta t(\hat{u}_{i,t} - l_{i,t})^2 + \rho_i(1 - z_{i,T})\right] \quad (18a)$$

$$\text{s.t.} \quad (\boldsymbol{u}_i, \boldsymbol{z}_i, \boldsymbol{s}_i, \boldsymbol{d}_i, \boldsymbol{r}_i) \in \mathcal{X}_i, \quad \forall i \in [N] \quad (18b)$$

$$\sum_{i=1}^{N} l_{i,t} = L_t, \quad \forall t \in [T] \quad (18c)$$

where $l_{i,t}$ is also a decision variable.

Consequently, the only substantial coupling among customers in the modified SOC-2 model (18) is the equality constraint (18c). Therefore, we can introduce the dual variable $\boldsymbol{\lambda} := (\lambda_t)_{t \in [T]}$ for the equality constraint (18c) and employ the *dual gradient algorithm* [31] to solve the modified SOC-2 model (18) in a distributed manner. The specific distributed solution method is presented as Algorithm 2. The implementation of Algorithm 2 needs the two-way communication of $\boldsymbol{\lambda}$ and $\boldsymbol{l}_i$ between the SA and every customer in each iteration. Due to the simple structure with only one equality coupling constraint (18c), Algorithm 2 can converge quickly with appropriate step sizes, which is verified by our simulations.

Similar to (17), we can derive the equivalent analytic form (21) for the expectation term in (19a):

$$\Delta t \sum_{t=1}^{T}\left\{(u_{i,t}^{\text{set}} - l_{i,t})^2 \right.$$
$$+ \left[2(u_{i,t}^{\text{set}} - l_{i,t})(u_{i,t} - u_{i,t}^{\text{set}}) + (u_{i,t} - u_{i,t}^{\text{set}})^2\right]$$
$$\left. \cdot \prod_{\tau=1}^{t-1} p_{\boldsymbol{\theta}_i}(\boldsymbol{w}_{i,\tau})\right\} - \rho_i \prod_{t=1}^{T} p_{\boldsymbol{\theta}_i}(\boldsymbol{w}_{i,t}). \quad (21)$$

By substituting the expectations terms in (16a) and (19a) with their analytic forms (17) and (21) respectively, the local AC control models (16) and (19) become deterministic nonconvex optimization problems. Given parameter $\boldsymbol{\theta}_i$, they can be solved efficiently via available nonlinear optimizer tools, such as the IPOPT solver [32]. For concise expression, we denote the

---

[3]The approximation is made by dropping the term $2\sum_{t=1}^{T}\sum_{i \neq j}\mathbb{E}(\hat{u}_{i,t} - l_{i,t})\mathbb{E}(\hat{u}_{j,t} - l_{j,t})$ from the expansion of (15a). This term is expected to be relatively small and thus neglectable.

Algorithm 2 Distributed Solution Algorithm

1: **Initialization:** Set initial dual value $(\lambda_t^0)_{t\in[T]}$, step size γ, convergence tolerance ϵ, and iteration count $k \leftarrow 0$.

2: **Parallel Optimization**: With the broadcast dual value $(\lambda_t^k)_{t\in[T]}$, each customer i solves the local AC control problem (19) and obtains the optimal solution $(\boldsymbol{u}_i^*, \boldsymbol{l}_i^*)$.

$$\min_{u_{i,t}, l_{i,t}} \mathbb{E}_{z_i}\left[\Delta t \sum_{t=1}^{T} (\hat{u}_{i,t} - l_{i,t})^2 + \rho_i(1 - z_{i,T})\right] + \sum_{t=1}^{T} \lambda_t^k l_{i,t} \tag{19a}$$

$$\text{s.t. } (\boldsymbol{u}_i, \boldsymbol{z}_i, \boldsymbol{s}_i, \boldsymbol{d}_i, \boldsymbol{r}_i) \in \mathcal{X}_i; \ l_{i,t} \geq 0, \ \forall t \in [T]. \tag{19b}$$

3: **Dual Variable Update**: Each customer i uploads $\boldsymbol{l}_i^*$ to the SA, and the SA updates the dual variable by

$$\lambda_t^{k+1} \leftarrow \lambda_t^k + \gamma \cdot \left(\sum_{i=1}^{N} l_{i,t}^* - L_t\right), \quad \forall t \in [T]. \tag{20}$$

4: **Convergence Check**: If $\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\| \leq \epsilon$, terminate the algorithm. Otherwise, let $k \leftarrow k + 1$ and go to Step 2.

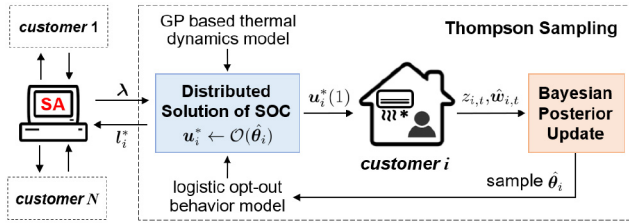

Fig. 2. Schematic of the distributed online DR control algorithm implemented on the SOC-2 problem. (For the SOC-1 problem, the communications of $\boldsymbol{\lambda}$ and $\boldsymbol{l}_i^*$ are not required.)

above distributed solution methods together with the optimizer tools as an oracle

$$\mathcal{O} : \boldsymbol{\theta}_i \to \boldsymbol{u}_i^*, \tag{22}$$

which generates optimal $\boldsymbol{u}_i^*$ with the input of parameter $\boldsymbol{\theta}_i$.

C. Distributed Online DR Control Algorithm

Based on the TS framework, we develop the distributed online DR control algorithm as Algorithm 3 to learn customer opt-out behaviors and optimally control the AC power in a DR event. Since this online algorithm is implemented distributedly, we present Algorithm 3 from the perspective of an individual customer $i \in [N]$. The practical implementation of Algorithm 3 is illustrated as Figure 2.

In Algorithm 3, the unknown customer behavior parameter $\boldsymbol{\theta}_i$ is treated as a random variable, and we construct a Gaussian prior distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ for it based on historical information. At each time $t \in [T]$ of a DR event, $\hat{\boldsymbol{\theta}}_i$ is randomly sampled from the distribution for decision-making. Two key techniques used in Algorithm 3 are explained as follows.

1) To utilize the latest information and take future time-slots into account, we employ the model predictive control (MPC) method in the optimization and action steps. Specifically, at

---

Algorithm 3 Distributed Online DR Control Algorithm

1: **Input:** Prior distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Receive the DR objective and associated parameters from the SA.

2: **for** $t = 1$ to T **do**

3:     **Sample** $\hat{\boldsymbol{\theta}}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

4:     **Optimization**. Solve the SOC model (14) or (18) distributedly for the remaining time horizon $\{t, \cdots, T\}$, and obtain the optimal AC power control scheme using the oracle $\boldsymbol{u}_i^* \leftarrow \mathcal{O}(\hat{\boldsymbol{\theta}}_i)$ (22).

5:     **Action**. Implement the first control action $\boldsymbol{u}_i^*(1)$, and observe the customer opt-out outcome $z_{i,t}$. Collect the environmental factors $\hat{\boldsymbol{w}}_{i,t}$ at time t.

6:     **Posterior Update**. Initialize variational parameter ξ_i by

$$\xi_i \leftarrow \sqrt{\hat{\boldsymbol{w}}_{i,t}^\top \boldsymbol{\Sigma}_i \hat{\boldsymbol{w}}_{i,t} + (\hat{\boldsymbol{w}}_{i,t}^\top \boldsymbol{\mu}_i)^2}. \tag{23}$$

    Iterate three times between the posterior update

$$\begin{cases} \hat{\boldsymbol{\Sigma}}_i^{-1} \leftarrow \boldsymbol{\Sigma}_i^{-1} + 2|\ell(\xi_i)|\hat{\boldsymbol{w}}_{i,t}\hat{\boldsymbol{w}}_{i,t}^\top, \\ \hat{\boldsymbol{\mu}}_i, \leftarrow \hat{\boldsymbol{\Sigma}}_i\left[\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i + (z_{i,t} - \frac{1}{2})\hat{\boldsymbol{w}}_{i,t}\right], \end{cases} \tag{24}$$

    where $\ell(\xi_i) := (1/2 - 1/(1 + e^{-\xi_i}))/2\xi_i$.

    and the ξ_i update

$$\xi_i \leftarrow \sqrt{\hat{\boldsymbol{w}}_{i,t}^\top \hat{\boldsymbol{\Sigma}}_i \hat{\boldsymbol{w}}_{i,t} + (\hat{\boldsymbol{w}}_{i,t}^\top \hat{\boldsymbol{\mu}}_i)^2}. \tag{25}$$

    Then set $\boldsymbol{\Sigma}_i \leftarrow \hat{\boldsymbol{\Sigma}}_i, \ \boldsymbol{\mu}_i \leftarrow \hat{\boldsymbol{\mu}}_i$.

7:     **Check Termination**. If $z_{i,t} = 0$, terminate the AC control and change the AC operation mode back to the original customized setting.

8: **end for**

---

each time t, it solves the SOC model (14) or (15) for the rest of the DR event to obtain the optimal AC control scheme $\boldsymbol{u}_i^*$, but only implements the first control action $\boldsymbol{u}_i^*(1)$. In addition, the latest predictions or estimations of the environmental factors, the updated thermal dynamics model, and recalculated baseline AC power $u_{i,t}^{\text{set}}$ can be adopted in the optimization step.

2) After observing the outcome pair $(z_{i,t}, \hat{\boldsymbol{w}}_{i,t})$, the *variational Bayesian inference approach* introduced in [33] is applied to obtain the posterior distribution on $\boldsymbol{\theta}_i$ with the update rules (23)-(25). It is well known that Bayesian inference for the logistic regression model (11) is an intrinsically hard problem [34], and the exact posterior $\mathcal{P}(\boldsymbol{\theta}_i|z_{i,t}, \hat{\boldsymbol{w}}_{i,t})$ is intractable to compute. Thus we use the variational approach [33] for efficient Bayesian inference, which provides an accurate Gaussian approximation to the exact posterior with a closed form (24). The scalar ξ_i in (23)-(25) is an intermediate parameter that affects the approximation accuracy, thus we alternate three times between the posterior update (24) and the ξ_i update (25), which leads to an optimal approximated posterior. See [33] for details.

D. Performance Measurement

For online learning problems, the notion of "regret" and its variants are standard metrics that are defined to measure the performance of online learning and decision algorithms [35].

Accordingly, we denote $\boldsymbol{\theta}^\star := (\boldsymbol{\theta}_i^\star)_{i\in[N]}$ as the underlying true customer behavior parameter that the LSEs do not know but aim to learn. Then the regret of the proposed online DR control algorithm at m-th DR event is defined as

$$\text{regret}(m) := C_m^{\boldsymbol{\theta}^*}\left(\boldsymbol{u}_m^{\text{online}}\right) - C_m^{\boldsymbol{\theta}^*}\left(\boldsymbol{u}_m^\star\right), \qquad (26)$$

where $C_m^{\boldsymbol{\theta}^*}(\cdot)$ denotes the objective function (i.e., (14a) in the SOC-1 model or (18a) in the modified SOC-2 model) under the true value $\boldsymbol{\theta}^\star$. $\boldsymbol{u}_m^{\text{online}}$ denotes the AC control scheme generated by the proposed online algorithm, while $\boldsymbol{u}_m^\star$ is the optimal AC control scheme that minimizes the objective $C_m^{\boldsymbol{\theta}^*}(\cdot)$. Thus regret($m$) in (26) is always non-negative and measures the performance distance between the proposed online algorithm and the underlying best control scheme.

To evaluate the overall learning performance, we further define the cumulative regret until M-th DR event as

$$\text{cumu\_regret}(M) := \sum_{m=1}^{M} \text{regret}(m), \qquad (27)$$

which is simply the sum of regret(m) over the first M DR events. Generally, a sublinear cumulative regret over M is desired, i.e., cumu_regret(M)$/M \to 0$ as $M \to +\infty$, since it indicates that regret(m) $\to 0$ as $m \to +\infty$. In other word, it means that the proposed online algorithm can eventually learn the customer behaviors well and make optimal AC control schemes as more and more DR events are experienced. We demonstrate the regret results of the proposed algorithm via numerical simulations in the next section.

V. NUMERICAL SIMULATIONS

In this section, we first test the performance of the linear thermal dynamics model and the GP model. Then, we implement the proposed distributed online DR control algorithm on the two SOC models.

A. Indoor Thermal Dynamics Prediction

In this part, we compare the thermal dynamics prediction performance of the linear model (9) and the GP model (10). Real customer metering data, including indoor temperature and AC power, from ThinkEco[4] are used for model training and testing. The outdoor temperature data are procured from an open-access meteorological database.[5] Specifically, we use the time series of data in consecutive 5 days with the resolution of 15 minutes to fit the two thermal dynamics models (9) and (10). The GPML toolbox [36] is applied to implement the GP model and optimize the hyperparameters. Then, the fitted models are tested on the time series of real data in the next 3 days for indoor temperature prediction.

The prediction results of one time step ahead (15 minutes) and three time steps ahead (45 minutes) are presented in Figure 3. The average prediction errors of the indoor temperature are $0.632°F$ and $0.737°F$ for the GP model in these
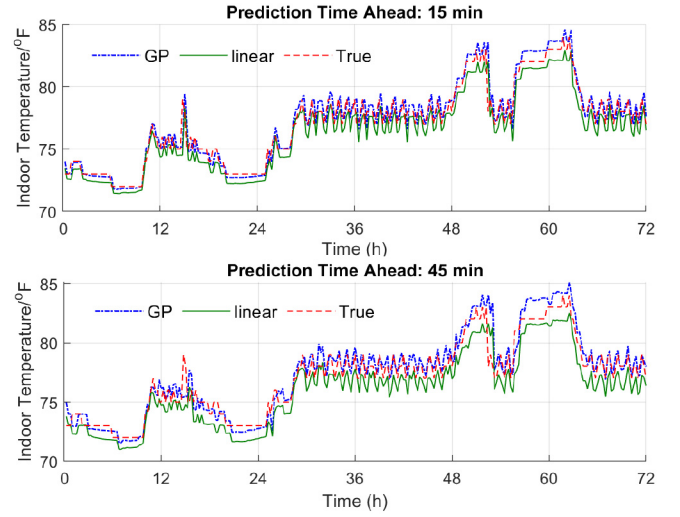

Fig. 3. Indoor temperature prediction results with Gaussian process and linear thermal dynamics models. (GP model: blue dotted curve; Linear model: green curve; True indoor temperature: red dashed curve.)

---

two cases, and $0.846°F$ and $1.16°F$ for the linear model. It is observed that both the GP model (10) and the linear model (9) work well in the thermal dynamics modeling, and the GP model achieves better prediction accuracy.

B. Learning and AC Control With SOC-1 Model

1) Simulation Configuration: Each DR event lasts for 3 hours with the AC control period $\Delta t = 15$ minutes, which implies a time length $T = 12$. The AC capacity and drift limit are set as $\bar{u}_i = 2\text{kW}$ and $\Delta u_i = 1\text{kW}$, and the AC setting temperature s_i^{set} is $72°F$. As defined in Section IV-D, we associate each customer $i \in [N]$ with a true behavior parameter $\boldsymbol{\theta}_i^\star \in \mathbb{R}^6$ to simulate the opt-out outcomes, whose values are randomly generated but satisfy several basic rules to be reasonable. For example, if no DR control is implemented (defaulted AC setting), the stay-in probability $p_{\boldsymbol{\theta}_i^\star}(\boldsymbol{w}_{i,t})$ should be very close to 1; if the indoor temperature $s_{i,t}$ reaches a high value such as $90°F$, the stay-in probability should be very close to 0. The considered environmental factors include indoor temperature $s_{i,t}$, accumulated thermal discomfort $d_{i,t}$, incentive credit $r_{i,t}$, outdoor temperature $s_{i,t}^{\text{out}}$, and time-varying electricity price, where the first three factors follow the dynamics (4)-(6) respectively, while the electricity price at each time is normalized and randomly generated from Unif(0, 1). Besides, IPOPT solver [32] is employed to solve the nonconvex optimal AC control models.

2) Control and Learning Performance: Since the aggregation of all local optimal AC control schemes is an optimal solution to the SOC-1 model (14), we simulate the AC control and learning for a single customer over sequential DR events. Given the true parameter $\boldsymbol{\theta}_i^\star$, the optimal AC power trajectory $\boldsymbol{u}_i^\star$ can be computed via solving the local control model (16). Figure 4 illustrates the simulation results associated with the optimal AC control scheme $\boldsymbol{u}_i^\star$. It is seen that the stay-in probability is maintained close to 1 by the AC control scheme, which tends to make customers comfortable and not opt out for the sake of long-term load reduction. Besides, there

---

[4]ThinkEco Inc. is a New York City-based energy efficiency and demand response solution company (http://www.thinkecoinc.com/).

[5]Iowa Environmental Mesonet [online]: https://mesonet.agron.iastate.edu/request/download.phtml?network=MA_ASOS.
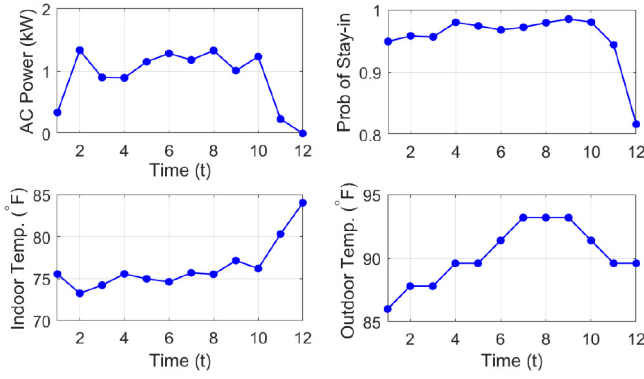
Fig. 4. Illustration of the optimal AC power control scheme $u_i^\star$, stay-in probability $p_{\theta_i^\star}(w_{i,t})$, indoor temperature $s_{i,t}$, outdoor temperature $s_{i,t}^{out}$ for a customer, which are computed via solving model (16) with known $\theta_i^\star$.
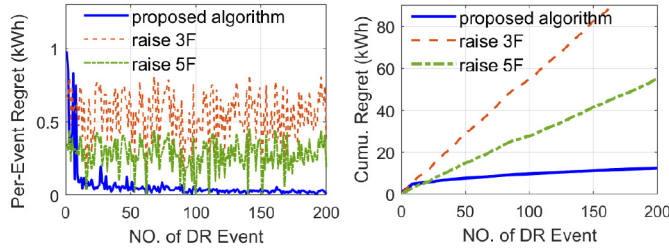

Fig. 5. The regret comparison between the proposed Algorithm 3 and two baseline schemes (raising the AC setting temperature by $3°F$ and $5°F$) for a typical customer under the SOC-1 model (14).
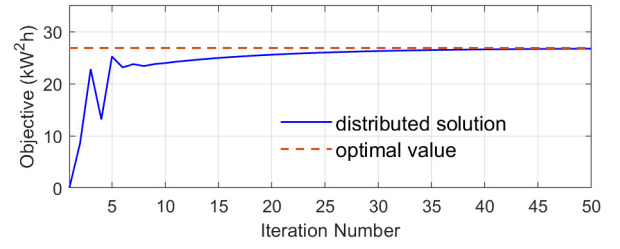

Fig. 6. Convergence results of the distributed algorithm (Algorithm 2) for solving the modified SOC-2 model (18).
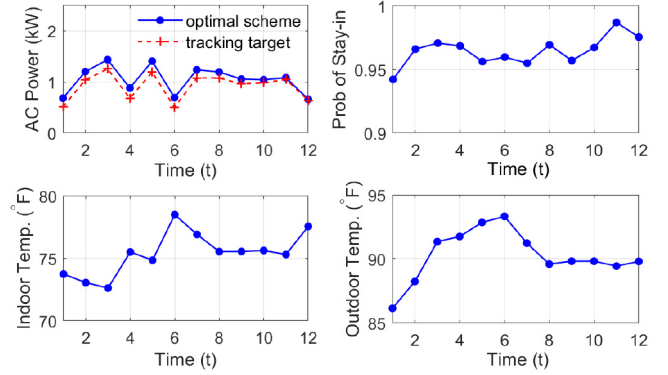

Fig. 7. Illustration of the optimal AC power control scheme $u_i^\star$ with the local tracking target $l_i^\star$, stay-in probability $p_{\theta_i^\star}(w_{i,t})$, indoor temperature $s_{i,t}$, outdoor temperature $s_{i,t}^{out}$ for a customer, which are computed via solving the modified SOC-2 model (18) with known $\theta_i^\star$.

is a drop of AC power at the end of the DR event ($t = 11, 12$), leading to increased indoor temperature and a decrease in the stay-in probability. Intuitively, that is because last-minute opt-out will not affect the DR objective much, and thus a radical AC power reduction is conducted. This effect can be mitigated by increasing the penalty coefficient ρ_i.

In practice, the true customer parameter $\theta_i^\star$ is unknown, and we implement the proposed online algorithm to learn customer behaviors and make AC control decisions in sequential DR events. We compare the performance of the proposed algorithm with two baseline schemes: raising the AC setting temperature by $3°F$ and $5°F$ respectively, which are typically used in real DR programs. The regret results for one customer over 200 DR events are shown as Figure 5. It is observed that the per-event regret(m) of the proposed algorithm decreases dramatically within the first tens of DR events, then converges to almost zero. As a result, the associated cumulative regret exhibits a clear sublinear trend, which verifies the learning efficiency of the proposed online algorithm. In contrast, the baseline schemes that simply raise the AC setting temperature without consideration of customer opt-out behaviors maintain high regret values. Besides, the average amount of load energy reduction of one customer in a DR event is 1.13 kWh for the proposed algorithm, which is higher than the baseline schemes with 0.6 kWh and 0.82 kWh load reduction, respectively.

C. Learning and AC Control With SOC-2 Model

We then perform the proposed online DR control algorithm on the SOC-2 model, where the simulation configurations are

the same as Section V-B. Since the SOC-2 problem involves the coordination among all customers, we set the customer number as $N = 500$, and the global tracking target L_t is randomly generated from Unif(450, 550)(kW).

1) Distributed Solution: We apply Algorithm 2 to solve the modified SOC-2 model (18) distributedly over 500 customers. A diminishing step size with $\gamma_k = \max(5 \times 10^{-4}/\sqrt{k}, 1 \times 10^{-4})$ is employed to speed up the convergence. For the case with given $\theta_i^\star$, the convergence results are shown as Figure 6. It is seen that the distributed algorithm converges to the optimal value within tens of iterations. We note that the convergence curve in Figure 6 and the associated AC power are just intermediate computational values, which are not executed in practice, while only the final converged values are regarded as the AC power control scheme and are implemented. Figure 7 illustrates the simulation results associated with the converged optimal AC control scheme, including the optimal AC power with the local tracking target, the stay-in probability, and the indoor/outdoor temperature profiles for one customer.

2) Learning Performance: In practice, the true parameter $\theta_i^\star$ is unknown, therefore we implement the proposed online algorithm to learn customer opt-out behaviors and make AC control decisions with the modified SOC-2 model (18). The regret results of the proposed algorithm for 200 DR events over 500 customers are presented in Figure 8. A rapidly decreasing regret and a sublinear cumulative regret are observed, which verify that the proposed algorithm can learn the customer behaviors well and generate efficient AC control decisions.
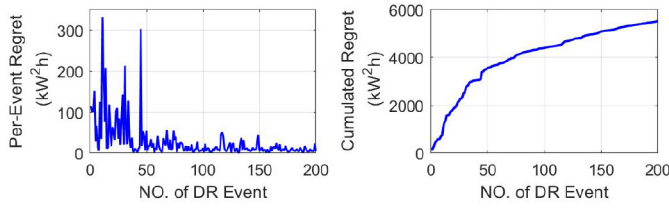
Fig. 8. The regret results of the proposed Algorithm 3 on the modified SOC-2 model (18) with 500 customers.

Besides, for the per-event regret curve in Figure 8, non-monotonic variations and occasional spikes are observed in the early learning stage (similar in Figure 5). That is because the proposed algorithm follows the TS framework and draws a random sample $\hat{\boldsymbol{\theta}}_i$ for decision-making at each time, which results in the non-monotonic variations with stochasticity. In addition, there is a small chance that the sample $\hat{\boldsymbol{\theta}}_i$ is quite different from the true value $\boldsymbol{\theta}_i^\star$, leading to large spikes in the regret curve. However, as the observed customer opt-out outcomes accumulate, the distribution on $\boldsymbol{\theta}_i$ gradually concentrates on the true value $\boldsymbol{\theta}_i^\star$, so that there are less large spikes in the later learning stage.

VI. CONCLUSION

In this paper, we propose a distributed online DR control algorithm to learn customer behaviors and regulate AC loads for incentive-based residential DR. Based on the Thompson sampling framework, the proposed algorithm consists of an online Bayesian learning step and an offline distributed optimization step. Two DR objectives, i.e., minimizing total AC loads and closely tracking a regulated power trajectory, are considered. The numerical simulations show that the distributed solution converges to the optimal value within tens of iterations, and the regret of learning reduces rapidly on average along with the implementation of DR events. Future works include 1) identify significant and effective environmental factors based on real user data; 2) conduct practical DR experiments using the proposed algorithm and analyze its practical performance.

APPENDIX
ANALYTIC DERIVATION FOR EXPECTATION IN (16a)

The expectation term in (16a) can be expanded as

$$\mathbb{E}_{z_i}\left[\sum_{t=1}^{T}\Delta t\left((u_{i,t}-u_{i,t}^{\text{set}})z_{i,t-1}\right)-\rho_i z_{i,T}\right]+\Delta t\sum_{t=1}^{T}u_{i,t}^{\text{set}}+\rho_i.$$

The status transition (3) implies that the customer opt-out time t_i^{opt} actually follows a geometric distribution with the Bernoulli probability $p_{i,t}:=p_{\boldsymbol{\theta}_i}(\boldsymbol{w}_{i,t})$ (11). Thus, we can enumerate all the possible realizations with different t_i^{opt} and the

probabilities:

$$\begin{cases}\Delta t\left(u_{i,1}-u_{i,1}^{\text{set}}\right), & \mathbb{P}\left(t_i^{\text{opt}}=1\right)=1-p_{i,1}\\ \cdots & \cdots\\ \Delta t\sum_{t=1}^{T}\left(u_{i,t}-u_{i,t}^{\text{set}}\right), & \mathbb{P}\left(t_i^{\text{opt}}=T\right)\\ & =\left(1-p_{i,T}\right)\prod_{t=1}^{T-1}p_{i,t}\\ \Delta t\sum_{t=1}^{T}\left(u_{i,t}-u_{i,t}^{\text{set}}\right)-\rho_i, & \mathbb{P}\left(t_i^{\text{opt}}=\text{not}\right)=\prod_{t=1}^{T}p_{i,t}\end{cases}$$

where "$t_i^{\text{opt}}=\text{not}$" means customer i does not opt out at any time. Then, the expectation term is computed analytically by summing up all these cases, which leads to (17) plus the constant term $\Delta t\sum_{t=1}^{T}u_{i,t}^{\text{set}}+\rho_i$.
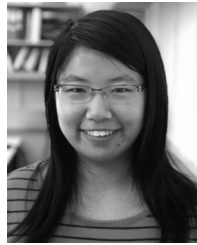
REFERENCES

[1] *2019 Assessment of Demand Response and Advanced Metering*, Federal Energy Regulatory Commission, Washington, DC, USA, Dec. 2019.

[2] *Benefit of Demand Response in Electricity Market and Recommendations for Achieving Them*, U.S. Dept. Energy, Washington, DC, USA, Feb. 2006.

[3] *Electric Power Annual*, U.S. Energy Inf. Admin., Washington, DC, USA, Oct. 2019.

[4] A. R. Jordehi, "Optimisation of demand response in electric power systems, a review," *Renew. Sustain. Energy Rev.*, vol. 103, pp. 308–319, Apr. 2019.

[5] M. Muratori and G. Rizzoni, "Residential demand response: Dynamic energy management and time-varying electricity pricing," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1108–1117, Mar. 2016.

[6] Q. Hu, F. Li, X. Fang, and L. Bai, "A framework of residential demand aggregation with financial incentives," *IEEE Trans. Smart Grid*, vol. 9, no. 1, pp. 497–505, Jan. 2018.

[7] Y. Li and N. Li, "Mechanism design for reliability in demand response with uncertainty," in *Proc. Amer. Control Conf. (ACC)*, Seattle, WA, USA, May 2017, pp. 3400–3405.

[8] X. Xu, C.-F. Chen, X. Zhu, and Q. Hu "Promoting acceptance of direct load control programs in the United States: Financial incentive versus control option," *Energy*, vol. 147, pp. 1278–1287, Mar. 2018.

[9] M. J. Fell, D. Shipworth, G. M. Huebner, and C. A. Elwell, "Public acceptability of domestic demand-side response in Great Britain: The role of automation and direct load control," *Energy Res. Soc. Sci.*, vol. 9, pp. 72–84, Sep. 2015.

[10] X. Xu, A. Maki, C.-F. Chen, B. Dong, and J. K. Day, "Investigating willingness to save energy and communication about energy use in the American workplace with the attitude-behavior-context model," *Energy Res. Soc. Sci.*, vol. 32, pp. 13–22, Oct. 2017.

[11] Q. Shi, C.-F. Chen, A. Mammoli, and F. Li, "Estimating the profile of incentive-based demand response (IBDR) by integrating technical models and social-behavioral factors," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 171–183, Jan. 2020.

[12] J. R. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: A review of algorithms and modeling techniques," *Appl. Energy*, vol. 235, pp. 1072–1089, Feb. 2019.

[13] X. Chen, G. Qu, Y. Tang, S. Low, and N. Li, "Reinforcement learning for decision-making and control in power systems: Tutorial, review, and vision," 2021. [Online]. Available: arXiv:2102.01168.

[14] Z. Wen, D. O'Neill, and H. Maei, "Optimal demand response using device-based reinforcement learning," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2312–2324, Sep. 2015.

[15] F. Alfaverh, M. Denaï, and Y. Sun, "Demand response strategy based on reinforcement learning and fuzzy reasoning for home energy management," *IEEE Access*, vol. 8, pp. 39310–39321, 2020.

[16] R. Lu, S. H. Hong, and M. Yu, "Demand response for home energy management using reinforcement learning and artificial neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6629–6639, Nov. 2019.

[17] H. Li, Z. Wan, and H. He, "Real-time residential demand response," *IEEE Trans. Smart Grid*, vol. 11, no. 5, pp. 4144–4154, Sep. 2020.

[18] F. Ruelens, B. J. Claessens, S. Vandael, B. De Schutter, R. Babuška, and R. Belmans, "Residential demand response of thermostatically controlled loads using batch reinforcement learning," *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2149–2159, Sep. 2017.

[19] R. Lu and S. H. Hong, "Incentive-based demand response for smart grid with reinforcement learning and deep neural network," *Appl. Energy*, vol. 236, pp. 937–949, Feb. 2019.

[20] X. Chen, Y. Nie, and N. Li, "Online residential demand response via contextual multi-armed bandits," *IEEE Contr. Syst. Lett.*, vol. 5, no. 2, pp. 433–438, Apr. 2021.

[21] Y. Li, Q. Hu, and N. Li, "A reliability-aware multi-armed bandit approach to learn and select users in demand response," *Automatica*, vol. 119, Sep. 2020, Art. no. 109015.

[22] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[23] S. Menard, *Applied Logistic Regression Analysis*, vol. 106. London, U.K.: Sage, 2002.

[24] D. J. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen, "A tutorial on Thompson sampling," *Found. Trends Mach. Learn.*, vol. 11, no. 1, pp. 1–96, 2018.

[25] A. Slivkins, "Introduction to multi-armed bandits," 2019. [Online]. Available: arXiv:1904.07272.

[26] D. Russo and B. V. Roy, "Learning to optimize via posterior sampling," *Math. Oper. Res.*, vol. 39, no. 4, pp. 1221–1243, Apr. 2014.

[27] J. Hall, C. Rasmussen, and J. Maciejowski, "Modelling and control of nonlinear systems using Gaussian processes with partial model information," in *Proc. 51st IEEE Conf. Decis. Control (CDC)*, Maui, HI, USA, Dec. 2012, pp. 5266–5271.

[28] D. P. Bertsekas, *Dynamic Programming and Optimal Control: Volume I*, 3rd ed. Belmont, MA, USA: Athena Sci., 2005.

[29] D. Liu, Y. Sun, Y. Qu, B. Li, and Y. Xu, "Analysis and accurate prediction of user's response behavior in incentive-based demand response," *IEEE Access*, vol. 7, pp. 3170–3180, 2019.

[30] N. Li, L. Chen, and S. H. Low, "Optimal demand response based on utility maximization in power networks," in *Proc. IEEE Power Energy Gen. Meeting*, Detroit, MI, USA, 2011, pp. 1–8.

[31] I. Necoara and V. Nedelcu, "On linear convergence of a distributed dual gradient algorithm for linearly constrained separable convex problems," *Automatica*, vol. 55, pp. 209–216, May 2015.

[32] *IPOPT Solver*. Accessed: 2021. [Online]. Available: https://coin-or.github.io/Ipopt/

[33] T. S. Jaakkola and M. I. Jordan, "A variational approach to Bayesian logistic regression models and their extensions," in *Proc. 6th Int. Workshop Artif. Intell. Stat.*, vol. 82, pp. 283–294, 1997.

[34] N. G. Polson, J. G. Scott, and J. Windle, "Bayesian inference for logistic models using Pólya–Gamma latent variables," *J. Amer. Stat. Assoc.*, vol. 108, no. 504, pp. 1339–1349, Dec. 2013.

[35] S. Shalev-Shwartz, "Online learning and online convex optimization," *Found. Trends Mach. Learn.*, vol. 4, no. 2, pp. 107–194, 2012.

[36] C. E. Rasmussen and H. Nickisch, "Gaussian processes for machine learning (GPML) toolbox," *J. Mach. Learn. Technol.*, vol. 11, pp. 3011–3015, Nov. 2010.

Yingying Li received the B.S. degree in mathematics and applied mathematics from the University of Science and Technology of China, Hefei, China, in 2015. She is currently pursuing the Ph.D. degree in applied mathematics with the John A. Paulson School of Engineering and Applied Sciences, Harvard University. Her research interest lies in online optimal control, online convex optimization, learning-based control with safety guarantees, multiarmed bandits, and reinforcement learning.

Jun Shimada received the Ph.D. degree in chemical physics from Harvard University and the Postdoctoral Research in theoretical biophysics with Columbia University and NYU. He is the Founder of ThinkEco, a start up focused on residential demand response. Prior to ThinkEco, he was a founding Team Member of Vitae Pharmaceuticals, where he was the Lead Scientist behind its computational drug discovery platform. His work has led to awards from various organizations, including the Alliance to Save Energy, International CES, and Powergrid International. A serial entrepreneur, he is always looking to apply rigorous theoretical concepts to the next big idea.

Xin Chen (Student Member, IEEE) received the double B.S. degrees in engineering physics and economics and the master's degree in electrical engineering from Tsinghua University, Beijing, China, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree in electrical engineering with Harvard University, USA. His research interests focus on distributed optimization, control, and learning of networked systems, with emphasis on power and energy systems. He received the Best Conference Paper Award in IEEE PES General Meeting 2016, and he was the Best Student Paper Award Finalist in the IEEE Conference on Control Technology and Applications 2018.

Na Li (Member, IEEE) received the bachelor's degree in mathematics from Zhejiang University in 2007, and the Ph.D. degree in control and dynamical systems from the California Institute of Technology in 2013. She is a Gordon McKay Professor of Electrical Engineering and Applied Mathematics with Harvard University. She was a Postdoctoral Associate with the Massachusetts Institute of Technology from 2013 to 2014. Her research lies in control, learning, and optimization of networked systems, including theory development, algorithm design, and applications to real-world cyber–physical societal system in particular power systems. She received the NSF Career Award in 2016, the AFSOR Young Investigator Award in 2017, the ONR Young Investigator Award in 2019, the Donald P. Eckman Award in 2019, the McDonald Mentoring Award in 2020, along with some other awards.