Online Residential Demand Response via Contextual Multi-Armed Bandits

Xin Chen, *Graduate Student Member, IEEE*, Yutong Nie, and Na Li, *Member, IEEE*

***Abstract*—Residential loads have great potential to enhance the efficiency and reliability of electricity systems via demand response (DR) programs. One major challenge in residential DR is how to learn and handle unknown and uncertain customer behaviors. In this letter, we consider the residential DR problem where the load service entity (LSE) aims to select an optimal subset of customers to optimize some DR performance, such as maximizing the expected load reduction with a financial budget or minimizing the expected squared deviation from a target reduction level. To learn the uncertain customer behaviors influenced by various time-varying environmental factors, we formulate the residential DR as a contextual multi-armed bandit (MAB) problem, and develop an online learning and selection (OLS) algorithm based on Thompson sampling to solve it. This algorithm takes the contextual information into consideration and is applicable to complicated DR settings. Numerical simulations are performed to demonstrate the learning effectiveness of the proposed algorithm.**

***Index Terms*—Residential demand response, online learning, multi-armed bandits, uncertainty.**

I. Introduction

With deepening penetration of renewable generation and growing peak load demands, electricity systems are inclined to confront a deficiency of reserve capacity for power supply-demand balance. Instead of deploying additional generators, demand response (DR) [1] is an economical and environmentally friendly alternative solution that calls for the change of flexible load demands to fit the needs of power supply. Since residential load takes up a significant share of the total electricity usage (e.g., about 38% in the U.S. [2]), it has huge potential to be exploited to facilitate power system operation. In residential DR programs [3], [4], the load service entities (LSEs) signal an upcoming DR event and recruit customers to participate with financial incentives, e.g., cash, coupon, raffle, rebate, etc. During the DR event, customers reduce their electricity consumption to earn the payment but are allowed to opt out. Since each recruitment comes with a cost, it is crucial for the LSEs to target right customers for DR participation.

However, in practice, the customer DR behaviors are highly uncertain and unknown despite the offered incentives [4], [5]. According to the investigations in [6], [7], the customer acceptances of DR load control are influenced by *individual preference* and *environmental factors*. The individual preference relates to customers' intrinsic socio-demographic characteristics, e.g., income, age, education, household size, attitude to energy saving, etc. The environmental factors refer to real-time externalities such as indoor temperature, offered incentives, electricity price, fatigue effect, weather conditions, etc. While LSEs barely have the access to customers' individual preferences or the knowledge how environmental factors affect their opt-out behaviors. Without considering the actual willingness, a blind customer selection scheme may lead to a high opt-out rate and inefficient load adjustment.

To address this issue, a natural idea is to learn unknown customer behaviors through interaction and observation. In particular, the multi-armed bandit (MAB) framework [8] can be employed to model the residential DR problem. MAB deals with uncertain decision-making problems where an agent selects actions (called "arms") sequentially and upon each selection observes a reward, while the agent is uncertain about how his selected actions affect the rewards. Through the observation of action-reward pairs, the agent can learn about the reward system and improve the selection strategies. Moreover, if the reward system is affected by some contexts, e.g., environmental factors and agent profiles, it is referred as contextual MAB [8]. Due to its simple and general structure, contextual MAB has been successfully applied in many fields, such as recommendation systems [9], clinical trials [10], Web advertisements [11], electric vehicle charging control [12], and etc.

Contextual MAB is capable to account for the influence of time-varying environmental factors on customer DR behaviors, which is missing or captured poorly in most existing work. This is in principle an effective approach to mitigate the unavoidable impacts incurred by the changes in the environment. For the existing literature, [13], [14] use reinforcement learning to learn the customer dissatisfaction on job delay, and automatically schedule the usage of household appliances under time-varying electricity price. In [15], [16],

Manuscript received March 17, 2020; revised May 19, 2020; accepted June 6, 2020. Date of publication June 17, 2020; date of current version July 2, 2020. This work was supported in part by NSF CAREER: under Grant ECCS-1553407, and in part by NSF EAGER: under Grant ECCS-1839632. Recommended by Senior Editor R. S. Smith. *(Corresponding author: Xin Chen.)*

Xin Chen and Na Li are with the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 USA (e-mail: chen_xin@g.harvard.edu; nali@seas.harvard.edu).

Yutong Nie is with the School of Mathematical Sciences, Zhejiang University, Hangzhou 310027, China (e-mail: ytnie@zju.edu.cn).

Digital Object Identifier 10.1109/LCSYS.2020.3003190

dynamical pricing schemes are designed based on the (contextual) MAB methods with the consideration of uncertain user power consumption to the price change. Based on ideas from crowdsourcing, [17] proposes an incentive-compatible MAB mechanism to design monetary offers to customers with unknown response characteristics to reduce power consumption. In [18]–[21], MAB and its variations, e.g., adversarial MAB and restless MAB, are applied to learn customer behaviors and unknown load parameters, while the upper-confidence-bound (UCB) algorithm, policy index, and other heuristic algorithms are used to select right customers for DR participation. In many of these research efforts [13]–[21], simplified DR problem formulations are generally used, and the influence of time-varying environmental factors is mostly neglected.

Contribution: In this letter, to deal with the uncertain customer behaviors and the environmental influence, we adopt the contextual MAB to model the residential DR as an online learning and decision-making problem. Specifically, this letter studies the DR problem where the LSE aims to select right customers to optimize some DR performance such as maximizing the expected load reduction with a financial budget or minimizing the expected squared deviation from a target reduction level. Based on the Thompson sampling (TS) framework [22], we develop an online learning and (customer) selection (OLS) algorithm to tackle this problem, where the logistic regression [23] is used to predict customer opt-out behaviors and the variational approximation approach [24] is employed for efficient Bayesian inference. With the decomposition into a Bayesian learning task and an offline optimization task, the proposed algorithm is applicable to practical DR applications with complicated settings. Moreover, theoretical performance guarantees on the proposed OLS algorithm are provided, which show that a sublinear Bayesian regret can be achieved. Lastly, the numerical simulations further demonstrate that harnessing the contextual information, including individual diversity and time-varying environmental factors, in the learning process improves the predictions on customer DR behaviors and leads to efficient customer selection schemes.

II. Problem Formulation

A. Residential DR Model

Consider the residential DR program with a system aggregator (SA) and N customers over a time horizon $[T] \coloneqq \{1, \dots, T\}$, where each time $t \in [T]$ corresponds to one DR event. As illustrated in Figure 1, there are two phases in a typical DR event [25]. Phase 0 denotes the preparation period, when the SA calls upon customers for load adjustment with incentives and selects a subset of participating customers under a certain budget. In phase 1, the selected customers adjust their electric usage while they can choose to opt out if feel unsatisfied. In the end, the SA pays the selected customers according to their contributions to the load adjustment. In the follows, we consider the case with a supply deficit and a need for load reduction, while the case with a supply surplus can be handled in the same way.

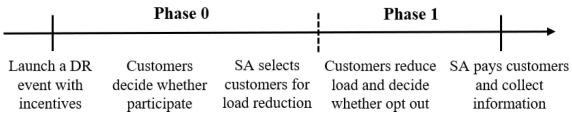

Fig. 1. Two phases of a residential DR event.

For customer $i \in [N] \coloneqq \{1, \dots, N\}$, let $d_{i,t}$ be the load reduction at t-th DR event, and $r_{i,t}$ be the credit paid to customer i for such load reduction. Denote $z_{i,t} \in \{1, 0\}$ as the binary variable indicating whether customer i stays in (equal 1) or opts out (equal 0) during t-th DR event if selected. Assume that $z_{i,t}$ is a random variable following Bernoulli distribution with $z_{i,t} \sim \text{Bern}(p_{i,t})$, which is independent across times and customers.[1]

This letter studies the customer selection strategies at phase 0 from the perspective of the SA. At t-th DR event, based on the reported $(d_{i,t}, r_{i,t})_{i\in[N]}$, the SA aims to select a subset of customers to achieve certain DR goals under the given budget b_t. Accordingly, the optimal customer selection (OCS) problem can be formulated as

$$\text{Obj.} \quad \max_{\mathcal{S}_t \subseteq [N]} \; g_t\big((z_{i,t})_{i\in[N]}, (d_{i,t})_{i\in[N]}, \mathcal{S}_t\big) \tag{1a}$$

$$\text{s.t.} \quad h_t\big((z_{i,t})_{i\in[N]}, (r_{i,t})_{i\in[N]}, \mathcal{S}_t\big) \le b_t, \;\; \mathcal{S}_t \in \Pi_t \tag{1b}$$

where g_t is the objective function describing the DR goal that the SA aims to optimize, and h_t represents the cost function that is constrained by budget b_t. $\mathcal{S}_t$ is the decision variable denoting the set of selected customers. Π_t is the feasible set of $\mathcal{S}_t$ that describes other physical constraints. For example, the network and power flow constraints can be captured by Π_t, which is elaborated in [29, Appendix A]. In practice, the set of available customers for each DR event may be different at phase 0 (some customers may not decide to participate). There are multiple ways for problem (1) to handle this issue, e.g., by setting $r_{i,t}$ as a very large value or using Π_t to model the time-varying available set of customers.

The OCS problem (1) is a general model. Depending on practical DR settings, functions g_t and h_t can take different forms. Two concrete examples are provided as follows.

Example 1: Model (2) maximizes the total expected load reduction, and constraint (2b) ensures that the total payment to customers does not exceed the given budget.

$$\text{Obj.} \quad \max_{\mathcal{S}_t \subseteq [N]} \; \mathbb{E}\Big(\sum_{i\in\mathcal{S}_t} d_{i,t} z_{i,t}\Big) = \sum_{i\in\mathcal{S}_t} d_{i,t} p_{i,t} \tag{2a}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{S}_t} r_{i,t} \le b_t. \tag{2b}$$

Example 2: Model (3) [18] aims to track a load reduction target D_t by minimizing the expected squared deviation in objective (3a). (3b) is a cardinality constraint on $\mathcal{S}_t$, which limits the number of selected customers by b_t. This can also

[1]The assumption of independence across times for each customer may be restrictive in practice, due to the fatigue effects and other potential dependencies. We cope with this limitation using the contextual model in next sub-section, which can capture the effects of time-varying factors.

be interpreted as the case with unit payment $r_{i,t} = 1$.

$$\text{Obj.} \quad \min_{\mathcal{S}_t \subseteq [N]} \mathbb{E}(\sum_{i \in \mathcal{S}_t} d_{i,t} z_{i,t} - D_t)^2 \tag{3a}$$

$$\text{s.t.} \quad |\mathcal{S}_t| = \sum_{i \in \mathcal{S}_t} 1 \le b_t. \tag{3b}$$

Since $z_{i,t} \sim \text{Bern}(p_{i,t})$ is assumed to be independent across customers, the objective in (3a) is equivalent to

$$\min_{\mathcal{S}_t \subseteq [N]} (\sum_{i \in \mathcal{S}_t} d_{i,t} p_{i,t} - D_t)^2 + \sum_{i \in \mathcal{S}_t} d_{i,t}^2 p_{i,t}(1 - p_{i,t}).$$

Remark: In our problem formulation, the offered credits $r_{i,t}$ and the budget b_t are given parameters. In practice, they can also be jointly designed with the learning process to achieve higher DR efficiency. This relates to the incentive mechanism design for DR programs [25], which is beyond the scope of this letter.

In the follows, Example 1 with the OCS model (2) is used for the illustration of algorithm design, while the proposed framework is clearly applicable to other application cases.

B. Contextual MAB Modelling

If the probability profiles $\boldsymbol{p}_t := (p_{i,t})_{i \in [N]}$ of customers are known, the OCS model (1) is purely an optimization problem. However, the SA barely has access to $\boldsymbol{p}_t$ in practice, which actually depict the customer opt-out behavior. Moreover, the probability profiles $\boldsymbol{p}_t$ are time-varying and influenced by various environmental factors. To address this uncertainty issue, the contextual MAB framework is leveraged to model the residential DR program as an online learning and decision-making problem. Specifically, each customer $i \in [N]$ is treated as an independent arm. At each time $t \in [T]$, the SA selects a set of customers $\mathcal{S}_t$, then observes the outcomes $(z_{i,t})_{i \in \mathcal{S}_t}$ that are generated from the distributions $\text{Bern}(p_{i,t})$, and receives the DR outcome $\sum_{i \in \mathcal{S}_t} d_{i,t} z_{i,t}$ as the reward in the MAB framework. Since the customer opt-out outcome $z_{i,t}$ is binary, the widely-used logistic regression method (4) is employed to model the unknown $p_{i,t}$:

$$p_{i,t} = \frac{\exp(\alpha_i + \boldsymbol{x}_{i,t}^\top \boldsymbol{\beta}_i)}{1 + \exp(\alpha_i + \boldsymbol{x}_{i,t}^\top \boldsymbol{\beta}_i)} \qquad \forall i \in [N], t \in [T] \tag{4}$$

where $\boldsymbol{x}_{i,t} \in \mathbb{R}^m$ is the feature vector that captures the environmental factors for customer i at t-th DR event. Each entry of $\boldsymbol{x}_{i,t}$ corresponds to a quantified factor, such as indoor/outdoor temperature (for air conditioner loads), weather condition (e.g., sunny or rainy), the offered credit $r_{i,t}$, real-time electricity price, the fatigue effect of being repeatedly selected, etc. $\boldsymbol{\beta}_i \in \mathbb{R}^m$ is the weight vector describing how customer i reacts to those factors, and α_i denotes his individual preference. Denote $\hat{\boldsymbol{x}}_{i,t} := (1, \boldsymbol{x}_{i,t})$ as the context vector and $\boldsymbol{\theta}_i := (\alpha_i, \boldsymbol{\beta}_i)$, then the linear term in (4) becomes $\hat{\boldsymbol{x}}_{i,t}^\top \boldsymbol{\theta}_i$, and the unknown parameter of each customer i is summarized by $\boldsymbol{\theta}_i$.

As a result, the sequential customer selection in residential DR is modelled as a contextual MAB problem. The SA aims to learn the unknown $(\boldsymbol{\theta}_i)_{i \in [N]}$ and improves the customer selection strategies. To this end, we propose the online learning and selection (OLS) algorithm to solve this contextual MAB problem efficiently in the next section.

Algorithm 1 Thompson Sampling Algorithm [22]

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

III. Algorithm Design

In this section, we first introduce the Thompson Sampling (TS) algorithm, the offline optimization method, and the Bayesian inference method. Then we assemble these methods to develop the OLS algorithm for residential DR.

A. Thompson Sampling Algorithm

Consider a classical T-times MAB problem where an agent selects an arm (action) a_t from the set $\mathcal{A}_t$ at each time t. After pulling the arm a_t, the agent observes an outcome u_t, which is randomly generated from a conditional probability distribution $\mathcal{P}_\theta(\cdot|a_t)$, and then obtains a reward $R_t = R(u_t)$ with known deterministic function $R(\cdot)$. The agent intends to maximize the total expected reward but is initially uncertain about the value of θ. TS algorithm [22] is a Bayesian learning framework that solves such MAB problems while effectively balancing exploration and exploitation.

As illustrated in Algorithm 1, TS algorithm represents the initial belief on θ using 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 based on the sample $\hat{\theta}$ (Step 4). After outcome u_t is observed, the Bayesian rule is applied to update the belief and obtain the posterior distribution of θ (Step 5). There are three key observations about the TS algorithm:

1) As outcomes accumulate, the predefined prior distribution will be washed out and the posterior converges to the true distribution or true value of θ.
2) The TS algorithm encourages exploration by the random sampling (Step 3). As posterior distribution gradually concentrates, less exploration and more exploitation will be performed, which strikes an effective balance.
3) The key advantage of TS algorithm is that the complex online problem is decomposed into a Bayesian learning task (Step 5) and a deterministic optimization task (Step 4) [26], and the optimization remains the original model formulation, which enables efficient solution.

Motivated by the last observation, in the follows, we describe the offline optimization method for Step 4 and the Bayesian update for Step 5 in Algorithm 1, respectively.

B. Offline Optimization Method

At each DR event, given the customer probability profiles $\boldsymbol{p}_t$, the OCS model (2) can be equivalently reformulated as the binary optimization problem (5):

$$\text{Obj.} \quad \max_{y_{i,t} \in \{0,1\}} \sum_{i=1}^{N} d_{i,t} p_{i,t} y_{i,t} \tag{5a}$$

$$\text{s.t.} \sum_{i=1}^{N} r_{i,t} y_{i,t} \le b_t \tag{5b}$$

where binary variable $y_{i,t} \in \{0, 1\}$ is introduced to indicate whether the SA selects customer i or not at t-th DR event.

The binary optimization model (5) can be solved efficiently using many available optimization solvers such as IBM CPLEX and Gurobi, which are employed as the offline solution tools. Let $\boldsymbol{y}_t^* := (y_{i,t}^*)_{i\in[N]}$ be the optimal solution of model (5). For concise expression, the optimizer tools are denoted as an offline oracle $\mathcal{O} : \boldsymbol{p}_t \to \boldsymbol{y}_t^*$.

C. Bayesian Inference for Logistic Model

For a simpler exposition, we abuse notations a little bit and discard subscripts i and t in this part as the Bayesian update rule is the same across customers and times. Under the TS framework, a prior distribution $\mathcal{P}(\boldsymbol{\theta})$ on the unknown parameter $\boldsymbol{\theta}$ is constructed. After the outcome $(\hat{\boldsymbol{x}}, z)$ is observed, the posterior distribution is calculated by the Bayesian law:

$$\mathcal{P}(\boldsymbol{\theta}|\hat{\boldsymbol{x}}, z) = \frac{\mathcal{P}(\boldsymbol{\theta})\mathcal{P}(z|\boldsymbol{\theta}, \hat{\boldsymbol{x}})}{\int_{\tilde{\boldsymbol{\theta}}} \mathcal{P}(\tilde{\boldsymbol{\theta}})\mathcal{P}(z|\tilde{\boldsymbol{\theta}}, \hat{\boldsymbol{x}})\, d\tilde{\boldsymbol{\theta}}} \tag{6}$$

and the logistic likelihood function $\mathcal{P}(z|\boldsymbol{\theta}, \hat{\boldsymbol{x}})$ is given by

$$\mathcal{P}(z|\boldsymbol{\theta}, \hat{\boldsymbol{x}}) = \phi\Big((2z-1)\hat{\boldsymbol{x}}^\top \boldsymbol{\theta}\Big) \tag{7}$$

where $\phi(x) := 1/(1+e^{-x})$ and (7) is equivalent to (4).

Due to the analytically inconvenient form of the likelihood function, Bayesian inference for the logistic regression model is recognized as an intrinsically hard problem [27], thus the exact posterior $\mathcal{P}(\boldsymbol{\theta}|\hat{\boldsymbol{x}}, z)$ (6) is intractable to compute. To address this issue, the concept of conjugate prior [28] is leveraged to obtain a closed-form expression for the posterior update. Specifically, let the prior be a Gaussian distribution with $\mathcal{P}(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the variational Bayesian inference approach [24] is employed to approximate the logistic likelihood function (7) with a Gaussian-like distribution. The fundamental tool at the heart of this approach is a lower bound approximation of (7):

$$\mathcal{P}(z|\boldsymbol{\theta}, \hat{\boldsymbol{x}}) \ge \underbrace{\phi(\xi)\exp\left[\frac{s-\xi}{2} + \ell(\xi)(s^2-\xi^2)\right]}_{:=\mathcal{P}(z|\boldsymbol{\theta}, \hat{\boldsymbol{x}}, \xi)} \tag{8}$$

where $s := (2z-1)\hat{\boldsymbol{x}}^\top\boldsymbol{\theta}$, $\ell(\xi) := (1/2 - \phi(\xi))/2\xi$, and ξ is the variational parameter.

The variational distribution $\mathcal{P}(z|\boldsymbol{\theta}, \hat{\boldsymbol{x}}, \xi)$ in (8) has a convenient property that it depends on $\boldsymbol{\theta}$ only quadratically in the exponent. As the prior is a Gaussian distribution, we use the Gaussian-like variational distribution $\mathcal{P}(z|\boldsymbol{\theta}, \hat{\boldsymbol{x}}, \xi)$ to approximate the logistic likelihood function $\mathcal{P}(z|\boldsymbol{\theta}, \hat{\boldsymbol{x}})$ in the Bayesian inference (6). As a result, the posterior is also a Gaussian distribution $\mathcal{P}(\boldsymbol{\theta}|\hat{\boldsymbol{x}}, z) \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ with the closed-form update rule (9). See [24] for the detailed derivation.

$$\hat{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{\Sigma}^{-1} + 2|\ell(\xi)|\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top \tag{9a}$$

$$\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\Sigma}}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + (z-\frac{1}{2})\hat{\boldsymbol{x}}\right] \tag{9b}$$

Algorithm 2 Online Learning and Selection Algorithm

1: **Input:** $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ of each customer $i \in [N]$.
2: **for** $t = 1$ to T **do**
3: Receive $\{d_{i,t}, r_{i,t}, \hat{\boldsymbol{x}}_{i,t}\}$ from each customer $i \in [N]$. The SA sets the budget parameter b_t.
4: **for** customer $i = 1$ to N (in parallel) **do**
5: Sample $\hat{\boldsymbol{\theta}}_{i,t} \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.
6: $p_{i,t} \leftarrow 1/\Big[1 + \exp(-\hat{\boldsymbol{x}}_{i,t}^\top \hat{\boldsymbol{\theta}}_{i,t})\Big]$.
7: **end for**
8: Solve the OCS problem (5) with oracle $\boldsymbol{y}_t \leftarrow \mathcal{O}(\boldsymbol{p}_t)$. Select those customers for DR event t with $y_{i,t} = 1$, and observe the responses $z_{i,t}$.
9: **for** customer $i \in [N]$ with $y_{i,t} = 1$ **do**
10: Initialize ξ_i by $\xi_i \leftarrow \sqrt{\hat{\boldsymbol{x}}_{i,t}^\top \boldsymbol{\Sigma}_i \hat{\boldsymbol{x}}_{i,t} + (\hat{\boldsymbol{x}}_{i,t}^\top \boldsymbol{\mu}_i)^2}$.
11: Iterate three times between the posterior update

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

and the ξ_i update

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

12: Set $\boldsymbol{\Sigma}_i \leftarrow \hat{\boldsymbol{\Sigma}}_i$, $\boldsymbol{\mu}_i \leftarrow \hat{\boldsymbol{\mu}}_i$.
13: **end for**
14: **end for**

Since the posterior covariance matrix $\hat{\boldsymbol{\Sigma}}$ depends on the variational parameter ξ, its value needs to be specified such that the lower bound approximation in (8) is optimized. The optimal ξ is achieved by maximizing the expected complete log-likelihood function $\mathbb{E}[\log \mathcal{P}(\boldsymbol{\theta})\mathcal{P}(z|\boldsymbol{\theta}, \hat{\boldsymbol{x}}, \xi)]$, where the expectation is taken over $\mathcal{P}(\boldsymbol{\theta}|\hat{\boldsymbol{x}}, z, \xi^{\text{old}})$, and this leads to a closed form solution:

$$\xi = \sqrt{\hat{\boldsymbol{x}}^\top \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{x}} + (\hat{\boldsymbol{x}}^\top \hat{\boldsymbol{\mu}})^2} \tag{10}$$

Alternating between the posterior update (9) and the ξ update (10) monotonically improves the posterior approximation. The convergence of this procedure is very fast, which generally only needs two or three iterations [24].

D. Online Learning and Selection Algorithm

For each customer $i \in [N]$, we construct a Gaussian prior $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ on the unknown $\boldsymbol{\theta}_i$ based on historical information. Using the TS framework, the online learning and selection (OLS) algorithm for residential DR is developed as Algorithm 2. From Step 4 to Step 7, it generates a random sample of $\hat{\boldsymbol{\theta}}_{i,t}$ and then calculates the probability $p_{i,t}$ with the context $\hat{\boldsymbol{x}}_{i,t}$ for each customer. With the obtained probability profiles $\boldsymbol{p}_t$, the SA determines the optimal selection of customers by solving the OCS model (5), when available optimizers can be used for solution. After observing the behavior outcome $z_{i,t}$ of each selected customer, Step 9 - Step 13 update the posterior on $\boldsymbol{\theta}_i$ using the variational Bayesian inference approach. In Step 11, the alternation between the posterior update and the ξ_i update is performed three times to obtain accurate posterior approximation [24].

The OLS algorithm inherits the merits of TS and decomposes the online DR problem into Bayesian learning and offline optimization. Since the optimization problem remains

the original form without being corrupted by the learning task, the OLS algorithm can be applied to practical DR problems with complicated objectives and constraints. Besides, the closed-form formula for posterior update enables convenient Bayesian inference on the unknown parameters.

IV. Performance Analysis

In this section, we provide the performance analysis for the proposed OLS algorithm and prove that it achieves a $O(\sqrt{T}\log T)$ Bayesian regret bound for the OCS problem (2) when exact Bayesian inference is used.

A. Main Result

Define $\hat{m} := m + 1$ as the dimension of $\boldsymbol{\theta}_i$. Let $\boldsymbol{\Theta} := (\boldsymbol{\theta}_i)_{i\in[N]} \in \mathbb{R}^{N\hat{m}}$ collect all the unknown customer parameters. Denote the reward received at time t as $R_t := \sum_{i\in\mathcal{S}_t} d_{i,t} z_{i,t}$ and define the reward function $f_t^{\boldsymbol{\Theta}}$ as

$$f_t^{\boldsymbol{\Theta}}(\mathcal{S}_t) = \mathbb{E}(R_t|\boldsymbol{\Theta}, \mathcal{S}_t) = \sum_{i\in\mathcal{S}_t} \frac{d_{i,t}}{1+\exp\left(-\hat{\boldsymbol{x}}_{i,t}^\top \boldsymbol{\theta}_i\right)} \tag{11}$$

To measure the performance of the online algorithm, we define the T-period regret as

$$\text{Regret}(T, \boldsymbol{\Theta}) = \sum_{t=1}^{T} \mathbb{E}\Big[f_t^{\boldsymbol{\Theta}}(\mathcal{S}_t^*) - f_t^{\boldsymbol{\Theta}}(\mathcal{S}_t) \mid \boldsymbol{\Theta}\Big] \tag{12}$$

where $\mathcal{S}_t^* \in \arg\max_{\mathcal{S}_t\in\mathcal{A}_t} f_t^{\boldsymbol{\Theta}}(\mathcal{S}_t)$ is the optimal solution of model (2) with known $\boldsymbol{\Theta}$, and $\mathcal{A}_t$ denotes the feasible region described by constraint (2b), while $\mathcal{S}_t$ denoted the customer selection decision made by the online algorithm. The expectation in (12) is taken over the randomness of $\mathcal{S}_t$. Generally, a sublinear regret over the time length T is desired, which indicates that the online algorithm can eventually learn the optimal solution, since $\text{Regret}(T, \boldsymbol{\Theta})/T \to 0$ as $T \to \infty$.

Since $\boldsymbol{\Theta}$ is treated as a random variable in TS algorithm, we further define the T-period Bayesian regret as

$$\text{BayRegret}(T) = \mathbb{E}_{\boldsymbol{\Theta}\sim\mathcal{P}(\boldsymbol{\Theta})}\big[\text{Regret}(T, \boldsymbol{\Theta})\big] \tag{13}$$

which is the expectation of $\text{Regret}(T, \boldsymbol{\Theta})$ taken over the prior distribution $\mathcal{P}(\boldsymbol{\Theta})$ of $\boldsymbol{\Theta}$. It can be shown that asymptotic bounds on Bayesian regret are essentially asymptotic bounds on regret with the same order. See [26] for more explanations on Bayesian regret.

We further make the following two assumptions. Assumption 1 of exact Bayesian inference is generally made for theoretic analysis of Bayesian regret. Note that in practice, performing exact Bayesian inference is usually intractable without conjugate prior, which is our case, thus approximation approaches or Gibbs sampler are employed to obtain posterior samples [22]. Regret analysis for MAB learning with inexact Bayesian inference remains an open question.

Assumption 1: Exact Bayesian inference is performed in the OLS algorithm, i.e., the exact posterior distribution of $\boldsymbol{\Theta}$ is obtained and used for sampling at each time $t \in [T]$.

Assumption 2: $\boldsymbol{\Theta}$ is bounded with $||\boldsymbol{\Theta}||_\infty \le L$. Further, without loss of generality, the context vectors are normalized such that $\|\hat{\boldsymbol{x}}_{i,t}\|_\infty \le 1$ for all $i \in [N]$ and $t \in [T]$.

We show that the proposed OLS algorithm can achieve a sublinear Bayesian regret with the order of $O(\sqrt{T}\log T)$ for problem (2), which is stated formally as Theorem 1.

Theorem 1: Under Assumption 1 and 2, the Bayesian regret bound of the OLS algorithm for the OCS problem (2) is

$$\text{BayRegret}(T) \le O\Big(\bar{D}\hat{m}N^{\frac{3}{2}} e^{\hat{m}L}\bar{d}/\underline{d} \cdot \sqrt{T}\log T\Big)$$

where $\bar{d} := \sup_{i,t} d_{i,t}$, $\underline{d} := \inf_{i,t} d_{i,t} > 0$, and $\bar{D}$ is defined in (14).

The theoretic results in [26, Sec. 7] are applied to prove Theorem 1 in the next subsection.

B. Proof Sketch of Theorem 1

We first show that our problem formulation satisfies the two assumptions imposed in [26, Sec. 7]. For [26, Assumption 1], the reward function (11) is bounded by

$$0 \le f_t^{\boldsymbol{\Theta}} \le \sup_{t\in[T]} \sum_{i\in[N]} d_{i,t} := \bar{D}, \quad \forall t \in [T] \tag{14}$$

In terms of [26, Assumption 2], we have the following lemma, whose proof is provided in [29, Appendix B] .

Lemma 1: For all $t \in [T]$, $R_t - f_t^{\boldsymbol{\Theta}}(\mathcal{S}_t)$ conditioned on $(\boldsymbol{\Theta}, \mathcal{S}_t)$ is $\frac{1}{2}\bar{D}$-sub-Gaussian.

Under Assumption 2, define the reward function class as

$$\mathcal{F}_t := \Big\{f_t^{\boldsymbol{\Theta}} \mid \boldsymbol{\Theta} \in \Psi\Big\}, \ t \in [T] \tag{15}$$

where $\Psi := \{\boldsymbol{\Theta} \in \mathbb{R}^{N\hat{m}} \mid ||\boldsymbol{\Theta}||_\infty \le L\}$. The reward function class $\mathcal{F}_t$ is time dependent due to the time-variant $d_{i,t}$ and $\hat{\boldsymbol{x}}_{i,t}$. By applying the result in [26, Proposition 10], we have the following Bayesian regret bound.

Lemma 2: Under Assumption 1 and 2, the Bayesian regret of the OLS algorithm for problem (2) is bounded by

$$\text{BayRegret}(T) \le \sup_{t\in[T]} \Bigg\{1 + \Big[\dim_E(\mathcal{F}_t, T^{-1}) + 1\Big]\bar{D} + 8\bar{D}\sqrt{\dim_E(\mathcal{F}_t, T^{-1})(1 + o(1) + \dim_K(\mathcal{F}_t))T\log T}\Bigg\} \tag{16}$$

In (16), $\dim_K(\mathcal{F}_t)$ is the Kolmogorov dimension defined by [26, Definition 1] and $\dim_E(\mathcal{F}_t, T^{-1})$ denotes the $1/T$-eluder dimension defined by [26, Definition 3] of function class $\mathcal{F}_t$. Intuitively, the Kolmogorov dimension is related to the measure of complexity to learn a function class, while the eluder dimension captures how effectively the unknown values can be inferred from the observed samples.

We can further bound $\dim_E(\mathcal{F}_t, T^{-1})$ and $\dim_K(\mathcal{F}_t)$ in (16) and achieve the final results in Theorem 1. Due to the page limit, the detailed proofs are omitted here and provided in the online version [29, Appendeix B] of this letter.

V. Numerical Simulations

In this section, numerical simulations are performed on the OCS problem (2) to test the OLS algorithm. Consider the residential DR with $N = 1000$ customers. The reduced

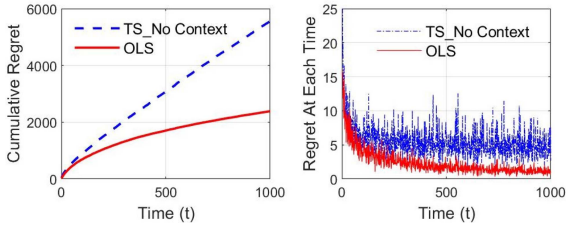

Fig. 2. Regret comparison between the OLS algorithm and the TS-based online learning algorithm without contexts.

load $d_{i,t}$ and the offered credit $r_{i,t}$ are randomly and independently generated from the uniform distribution Unif$[0, 1]$ for each customer, which are fixed for different time steps. At each time t, the SA randomly samples a budget b_t from Unif$[200, 300]$. Set the number of contextual features as $m = 9$, and let all customers share a same feature vector $\boldsymbol{x}_t \in \mathbb{R}^9$ at time t, whose elements are randomly generated from Unif$[0, 4]$. Assume that there exists an underlying ground truth $\boldsymbol{\theta}_i^* \in \mathbb{R}^{10}$ associated with each customer, which is generated from Unif$[-0.5, 0.5]$. In the OLS algorithm, we assign a Gaussian prior $\mathcal{N}(\boldsymbol{\theta}_i^* + 0.3\boldsymbol{u}_i, 0.09\boldsymbol{I})$ for each customer, where each element of $\boldsymbol{u}_i$ is randomly generated from Unif$[-1, 1]$ and $\boldsymbol{I}$ denotes the identity matrix.

To demonstrate the learning performance and exhibit the necessity of contextual information, we compare the OLS algorithm with a TS-based learning algorithm without context. Specifically, this "no-context" TS algorithm is derived from the Bernoulli bandit case with Beta prior distribution. See [22, Algorithm 2] for details. The simulation results are shown as Figure 2. The OLS algorithm achieves a sublinear cumulative regret and its regret at each time gradually decreases to zero. In contrast, due to neglecting contextual factors, the "no-context" TS algorithm does not learn user behaviors well and maintains high regret at each time.

VI. Conclusion

In this letter, the contextual MAB method is employed to model the customer selection problem in residential DR, considering the uncertain customer behaviors and the influence of contextual factors. Based on TS framework, the OLS algorithm is developed to learn customer behaviors and select appropriate customers for load reduction with the balance between exploration and exploitation. The simulation results demonstrate the necessity to consider the contextual factors and the learning effectiveness of the proposed algorithm. For future work, there are two attempt directions: 1) develop the optimal real-time control schemes for load devices during Phase 1 of the DR event, considering physical system dynamics; 2) study how to design the incentive mechanisms that optimize the credits and budget to achieve higher DR efficiency, together with the learning process.

References

[1] *Benefit of Demand Response in Electricity Market and Recommendations for Achieving Them*, U.S. Dept. Energy, Washington, DC, USA, Feb. 2006.
[2] *Electric Power Annual*, U.S. Energy Inf. Admin., Washington, DC, USA, Oct. 2019.
[3] H. Zhong, L. Xie, and Q. Xia, "Coupon incentive-based demand response: Theory and case study," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1266–1276, May 2013.
[4] 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.
[5] *Cost-Effectiveness of Electric Demand Response for Residential End-Uses: Final Report Prepared for National Grid*, Navigant Consult., Inc., Burlington, MA, USA, Apr. 2019.
[6] X. Xu, C. 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.
[7] 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.
[8] A. Slivkins, "Introduction to multi-armed bandits," 2019. [Online]. Available: arXiv:1904.07272.
[9] Y. Liu, Y. Xiao, Q. Wu, C. Miao, and J. Zhang, "Bandit learning for diversified interactive recommendation," 2019. [Online]. Available: arXiv: 1907.01647.
[10] S. S. Villar, J. Bowden, and J. Wason, "Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges," *Stat. Sci. Rev. J. Inst. Math. Stat.*, vol. 30, no. 2, pp. 199–215, May 2016.
[11] D. Chakrabarti, R. Kumar, F. Radlinski, and E. Upfal, "Mortal multi-armed bandits," in *Advances in Neural Information Processing Systems*. La Jolla, CA, USA: Curran Assoc., Inc, 2009, pp. 273–280.
[12] C. Römer, J. Hiry, C. Kittl, and T. Leibig, and C. Rehtanz, "Charging control of electric vehicles using contextual bandits considering the electrical distribution grid," 2019. [Online]. Available: arXiv:1905.01163.
[13] D. O'Neill, M. Levorato, A. Goldsmith, and U. Mitra, "Residential demand response using reinforcement learning," in *Proc. 1st IEEE Int. Conf. Smart Grid Commun.*, Gaithersburg, MD, USA, 2010, pp. 409–414.
[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] M. Brégère, P. Gaillard, Y. Goude, G. Stoltz, "Target tracking for contextual bandits: Application to demand side management," 2019. [Online]. Available: arXiv:1901.09532.
[16] A. Moradipari, C. Silva, and M. Alizadeh, "Learning to dynamically price electricity demand based on multi-armed bandits," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Anaheim, CA, USA, 2018, pp. 917–921.
[17] S. Jain, B. Narayanaswamy, and Y. Narahari, "A multiarmed bandit incentive mechanism for crowdsourcing demand response in smart grids," in *Proc. 28th AAAI Conf. Artif. Intell.*, Jun. 2014, pp. 721–727.
[18] Y. Li, Q. Hu, and N. Li, "Learning and selecting the right customers for reliability: A multi-armed bandit approach," in *Proc. IEEE Conf. Decis. Control (CDC)*, Miami Beach, FL, USA, Dec. 2018, pp. 4869–4874.
[19] A. Mohamed, A. Lesage-Landry, and J. A. Taylor, "Dispatching thermostatically controlled loads for frequency regulation using adversarial multi-armed bandits," in *Proc. IEEE Elect. Power Energy Conf. (EPEC)*, Saskatoon, SK, Canada, 2017, pp. 1–6.
[20] Q. Wang, M. Liu, and J. L. Mathieu, "Adaptive demand response: Online learning of restless and controlled bandits," in *Proc. IEEE Int. Conf. Smart Grid Commun.*, Venice, Italy, 2014, pp. 752–757.
[21] J. A. Taylor and J. L. Mathieu, "Index policies for demand response," *IEEE Trans. Power Syst.*, vol. 29, no. 3, pp. 1287–1295, May 2014.
[22] 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.
[23] S. Menard, *Applied Logistic Regression Analysis*, vol. 106. Thousand Oaks, CA, USA: Sage, 2002.
[24] 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, 1997, p. 4.
[25] H. Ma , V. Robu, N. Li, and D. C. Parkes, "Incentivizing reliability in demand-side response," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 352–358.
[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] 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.
[28] P. Diaconis and D. Ylvisaker, "Conjugate priors for exponential families," *Ann. Stat.*, vol. 7, no. 2, pp. 269–281, 1979.
[29] X. Chen, Y. Nie, and N. Li, "Online residential demand response via contextual multi-armed bandits," 2020. [Online]. Available: arXiv:2003.03627.