LEARNING TO DYNAMICALLY PRICE ELECTRICITY DEMAND BASED ON MULTI-ARMED BANDITS

Ahmadreza Moradipari, Cody Silva, Mahnoosh Alizadeh
University of California Santa Barbara

ABSTRACT

We consider the electricity price design problem faced by an aggregator running a real-time pricing program to shape the demand of its population of customers. To capture the effects of the stochastic and unknown nature of the load's price response structure, we adopt a multi-armed bandit framework and propose a Thompson Sampling based algorithm to minimize the aggregator's regret from running the real-time pricing program given exogenously changing grid conditions. We provide a discussion on regret bounds for our algorithm.

1. INTRODUCTION

Real-time retail electricity pricing (RTP) is a market-based framework to shape electricity demand through hourly varying prices. There are attractive features that make RTP popular, including decentralized implementation, little communication needs, and relatively fair resource allocation with little information. Yet, setting aside a few pilot programs that pass on wholesale prices to retail customers, RTP algorithms have not been adopted so far to shape residential and commercial loads. One important reason for this is a lack of knowledge on how customers respond to price signals.

It is not ideal to consider the problem of learning the response of a population of customers to price signals as a complete "black box problem". There are many reasons for this, including 1) the existence of random elements and exogenous parameters that lead to variability in the daily behavior of the load; 2) the variability of the control objective on a daily basis (e.g., due to randomness in renewable generation outputs); 3) complex inter-temporal correlations in load; and 4) the small size of the set of observations that one can gather (there are only 365 days in a year). In this paper, we exploit two factors that can reduce the complexity of this problem. Specifically, we first consider that load flexibility exhibits a lower-dimensional structure that we exploit in our framework. Second, we adopt statistical models that incorporate prior knowledge of how customers modify their usage patterns of individual appliances in response to price signals from behavioral studies [1]. Bayesian reinforcement learning and adaptive control methods provide the machinery to exploit such prior knowledge in dynamic pricing, and have been studied extensively in systems where the uncertain nature of human decision-making behavior affects system performance [2].

Specifically, we adopt a multi-armed bandit (MAB) framework to minimize the operational costs of an electricity aggregator running an RTP program [3–5]. We assume that prices are posted at the beginning of each day, and the cost incurred every day is a non-linear function of the load response and exogenously varying parameters such as renewable generation outputs or grid conditions. We propose a Thompson sampling based algorithm to minimize the aggregator's regret from not knowing the true form of the customers' price response [6–8]. Lastly, we provide a discussion on regret performance that considers the effect of the exogenously varying parameters on the performance of the algorithm [9, 10].

In the context of demand response (DR), a number of papers have considered online pricing methods [11]. For example, in [12] and [13], the authors consider a linear regression model to estimate the price response. In [14], the authors apply a combinatorial MAB framework to choose the right customers to target in a DR program in order to maximize grid reliability and propose a UCB based algorithm. In [15, 16], the authors propose a perturbed myopic policy for price design based on a least square estimator of the unknown demand parameters and discuss its regret performance. The authors in [17] design an online learning algorithm for price design, referred to as piecewise linear stochastic approximation. The authors in [18] use an online learning framework to estimate and control the load flexibility of a population of air conditioners and separate it from other loads. The authors in [19] use a MAB framework to select the right consumers to target for load reduction signals. Our paper focuses on the daily price design problem of an aggregator performing RTP and varies in terms of both load modeling and learning approach from all the above papers. The main advantage of Thompson sampling is its straightforward algorithmic implementation and the fact that it generalizes to more complex reward models, which are common in practical applications such as electricity pricing. In contract, UCB based approaches cannot be extended easily beyond generalized linear models, which are not generally sufficient for demand response optimization.

2. PROBLEM SETTING

Let us first present how we mathematically capture the price response of a population of electricity customers.

2.1. Load Flexibility Model

We adopt a general model of load flexibility that was first proposed in [20]. Specifically, we assume that the load flexibility of electric appliances can belong to a finite number of clusters $i \in \mathcal{I}$. Appliances in each cluster require approximately the same amount of energy and exhibit similar constraints and flexibility in consuming this energy. Accordingly, we can associate a set of feasible electricity consumption patterns (a.k.a, load profiles) $\mathcal{L}_i$ to each cluster. Any load profile $L_i(\tau) \in \mathcal{L}_i$ would satisfy the needs of an appliance in cluster i. For example, consider the cluster that represents electric vehicles (EVs) requiring a total of $E^i = 20$kWhs of charge between the hours $\tau_1^i = 6$pm and $\tau_2^i = 8$am. Assume charging can happen at a maximum rate of ρ. The set $\mathcal{L}_i$ of feasible hourly load profiles is given by:

$$\mathcal{L}_i = \{L_i(\tau) | \sum_{\tau=\tau_1^i}^{\tau_2^i} L_i(\tau) = E^i, 0 \leq L_i(\tau) \leq \rho\} \quad (2.1)$$

For a full discussion on characterizing the sets $\mathcal{L}_i$ for different types of flexible appliances, we refer the reader to [20].

Now, we consider a population of customers that own and operate a heterogeneous group of electric appliances and have different levels of flexibility in how their energy needs are satisfied. Mathematically, the total flexibility of the electricity demand of the population of customers can be characterized as a function of how many appliances belong to each cluster. If the number of appliances in each cluster is denoted as a_i, the set of load shapes $\mathcal{L}$ that can serve our population of customers can be written as:

$$\mathcal{L} = \sum_{i \in \mathcal{I}} a_i \mathcal{L}_i, \quad (2.2)$$

where the summation and scalar multiplication operations are defined in the sense of Minkowski addition[1].

Next, we discuss how a specific load shape $L(t)$ will emerge in response to a price signal $\mathbf{p} = [p(\tau)]_{\tau \in \Gamma}$.

2.2. Price Response

Dynamic pricing can have two effects on the customers demand for electricity, which we detail next.

[1]For two sets A and B defined on a finite dimensional Euclidean space, the Minkowski sum is defined as:

$$A + B = \{\mathbf{a} + \mathbf{b} \,|\, \mathbf{a} \in A,\ \mathbf{b} \in B\}. \quad (2.3)$$

1) Automated per cluster response: For appliances in cluster i, we assume all customers will choose $L_i^\star \in \mathcal{L}_i$ such that

$$L_i^\star(\tau, \mathbf{p}) = \text{argmin}_{L_i(\tau) \in \mathcal{L}_i} \sum_{\tau \in \Gamma} p(\tau) L_i(\tau). \quad (2.4)$$

This is a reasonable assumption given the automated nature of price response enabled through home energy management systems once the customer's preferences are set, e.g., the charge amount and the deadline to charge for an EV are specified.
2) Preference adjustment: The number of appliances in each cluster is driven by the economic price response of the customers that own and operate them. Hence, in response to dynamically varying prices, the customers can respond by adjusting their preferences, i.e., the number of appliances in each cluster becomes a function $a_i(\mathbf{p})$ of the posted price $\mathbf{p}$.

Together, the above two effects define the response of a population of electric appliances to a price signal $\mathbf{p}$ as:

$$L^\star(\tau, \mathbf{p}) = \sum_{i \in \mathcal{I}} a_i(\mathbf{p}) L_i^\star(\tau, \mathbf{p}). \quad (2.5)$$

Hence, given a set of feasible price signals $\mathbf{p} \in \mathcal{P}$ and a full characterization of the load response, i.e., knowing $a_i(\mathbf{p})$ and $\mathcal{L}_i$, it is clear that one can pick the price $\mathbf{p}^\star$ that shapes the demand according to a certain cost minimizing objective. However, in reality, the per cluster price response variables $a_i(\mathbf{p})$ are 1) random variables, i.e., the same price will not always elicit the same exact response from the customers. There is a certain level of randomness involved in the price response; 2) unknown to the aggregator, i.e., the aggregator does not know the structure of the underlying model that drives the price response $a_i(\mathbf{p})$; and 3) unobservable, i.e., the aggregator cannot see the disaggregated response of each cluster. Only the aggregate load $L^\star(t, \mathbf{p})$ in response to a posted price $\mathbf{p}$ would be observable to the aggregator.

2.3. The Price Design Objective

On day t, the aggregator's cost would depend on the load shape $\mathbf{L}_t^\star(\mathbf{p}_t) = [L_t^\star(\tau, \mathbf{p}_t)]_{\tau \in \Gamma}$ observed in response to a posted price $\mathbf{p}_t$, as well as an exogenous and random parameter vector $\mathbf{d}_t$ that is observable before the pricing decisions are made. The $\mathbf{d}_t$'s can mirror dynamically changing renewable generation outputs or grid conditions. For example, $\mathbf{d}_t$ can capture the target load profile on day t. The $\mathbf{d}_t$'s are i.i.d drawn from a distribution defined on a finite sample set $\mathcal{D}$, and each $\mathbf{d}_t$ occurs with a probability bounded away from zero, i.e., $\mathbb{P}(\mathbf{d}) > \varepsilon > \mathbf{0}, \forall \mathbf{d} \in \mathcal{D}$.

We allow for a nonlinear but fixed and known function $g(\mathbf{L}_t^\star(\mathbf{p}_t), \mathbf{d}_t)$ to represent the aggregator's cost on day t. Given the knowledge of the true model for the $a_i(\mathbf{p})$'s, the aggregator can choose the price $\mathbf{p}_t^\star$ that minimizes its daily cost $g(\mathbf{L}_t^\star(\mathbf{p}_t), \mathbf{d}_t)$. However, the true model of the per cluster sensitivities $a_i(\mathbf{p})$ is not available to the aggregator. Hence, the main question in this paper is, how can the aggregator

choose a series of prices $\mathbf{p}_t, t = 1, \ldots, T$ to minimize its regret from not knowing the true model of the $a_i(\mathbf{p})$'s over a horizon of T days, where regret is defined as:

$$R = \sum_{t=1}^{T} g(\mathbf{L}_t^\star(\mathbf{p}_t), \mathbf{d}_t) - \sum_{t=1}^{T} g(\mathbf{L}_t^\star(\mathbf{p}_t^\star), \mathbf{d}_t). \quad (2.6)$$

3. DYNAMIC PRICING USING MULTI-ARMED BANDITS

3.1. Bandit Setup

We consider the sensitivities $a_i(\mathbf{p})$ as random variables with parameterized distributions based on the posted price $\mathbf{p}$ and an unknown but constant parameter vector $\boldsymbol{\theta}^\star \in \Theta$ (which represents the *true model* for the customers' price response). A MAB setup allows us to capture the exploitation-exploration trade-off that emerges when learning the reliance of the stochastic behavior of the users' price response on the unknown parameter vector $\boldsymbol{\theta}^\star$ while trying to minimize costs. We propose a Thompson Sampling (TS) algorithm for the MAB-based dynamic pricing problem next.

3.2. A Thompson Sampling (TS) Based Approach

Thompson Sampling assumes that there is a prior distribution π available on the unknown parameters $\boldsymbol{\theta} \in \Theta$, with a non-zero probability associated with the true parameter $\boldsymbol{\theta}^\star$. On each day t, the algorithm makes a random draw $\boldsymbol{\theta}_t$ from the prior, and then acts optimally on choosing the electricity price $\mathbf{p}_t$ that minimizes expected costs, i.e., $\mathbb{E}\big[g(\mathbf{L}^\star(\mathbf{p}_t), \mathbf{d}_t))|\boldsymbol{\theta} = \boldsymbol{\theta}_t\big]$, conditional on that draw. After observing the load in response to this posted price, it performs a Bayesian update on the probability distribution π based on the new observation.

Denote $l(Y; \mathbf{p}, \boldsymbol{\theta}) = \mathbb{P}_\theta(\mathbf{L}^\star(\mathbf{p}_t) = Y|\mathbf{p}_t = \mathbf{p})$ as the likelihood of observing a load profile $\mathbf{L}^\star(\mathbf{p})$, upon posting a price $\mathbf{p}$ when the true parameter is $\boldsymbol{\theta}$. For the price design problem as defined above, our TS based algorithm is given by Alg. 1.

3.3. A Discussion on Regret Performance

Our regret analysis is inspired by the results in [9] for TS with nonlinear cost functions. We additionally analyze the effects of the exogenous parameters $\mathbf{d}_t$ on the algorithm's regret bounds. The analysis provides a bound on the total number of sub-optimal prices posted by the algorithm up to time T. Let $\mathbf{p}^{\star^\mathbf{d}}$ denote the optimal price posted when the model $\boldsymbol{\theta}_t$ drawn is equal to the true model $\boldsymbol{\theta}^\star$ and the target profile $\mathbf{d}$ is observed. Any price $\mathbf{p} \neq \mathbf{p}^{\star^\mathbf{d}}$ when the target profile $\mathbf{d}$ is observed is considered a suboptimal price. Before stating our results, we briefly explain how the posterior updates affect the performance of TS. When price $\mathbf{p}$ is posted at day t, the prior

Algorithm 1 Thompson Sampling

Input: Parameter space Θ, set of prices $\mathcal{P}$, output space $\mathcal{Y}$, likelihood $l(Y; \mathbf{p}, \boldsymbol{\theta})$.
Parameter: Distribution π over Θ.
Initialization: Set $\pi = \pi_0$.

for each day $t = 1, 2, \ldots T$

1. Draw $\boldsymbol{\theta}_t \in \Theta$ according to the distribution π_{t-1}.
2. Observe the exogenous parameter vector $\mathbf{d}_t$.
3. Post the optimal price

$$\mathbf{p}_t = \operatorname*{argmin}_{\mathbf{p}} \mathbb{E}\big[g(\mathbf{L}^\star(\mathbf{p}), \mathbf{d}_t))|\boldsymbol{\theta} = \boldsymbol{\theta}_t\big] \quad (3.1)$$

4. Observe $Y_t = \mathbf{L}^\star(\mathbf{p}_t)$.
5. (Posterior Update) set distribution π_t over Θ to

$$\forall S \subseteq \Theta : \pi_t(S) = \frac{\int_S l(Y_t; \mathbf{p}_t, \boldsymbol{\theta})\pi_{t-1}(d\boldsymbol{\theta})}{\int_\Theta l(Y_t; \mathbf{p}_t, \boldsymbol{\theta})\pi_{t-1}(d\boldsymbol{\theta})} \quad (3.2)$$

density is updated as

$$\pi_t(d\boldsymbol{\theta}) \propto \exp\left(-\log \frac{l(Y_t; \mathbf{p}, \boldsymbol{\theta}^\star)}{l(Y_t; \mathbf{p}, \boldsymbol{\theta})}\right) \pi_{t-1}(d\boldsymbol{\theta}). \quad (3.3)$$

Denote by $D(\boldsymbol{\theta}^\star_\mathbf{p}||\boldsymbol{\theta}_\mathbf{p})$ the marginal Kullback-Leibler divergence between the distribution $\{l(Y; \mathbf{p}, \boldsymbol{\theta}^\star) : Y \in \mathcal{Y}\}$ and $\{l(Y; \mathbf{p}, \boldsymbol{\theta}) : Y \in \mathcal{Y}\}$. As in [9], we can approximately write (3.3) as:

$$\pi_t(d\boldsymbol{\theta}) \propto \exp\left(-\sum_{p \in \mathcal{P}} N_t(\mathbf{p}) D(\boldsymbol{\theta}^\star_\mathbf{p}||\boldsymbol{\theta}_\mathbf{p})\right) \pi_{t-1}(d\boldsymbol{\theta}), \quad (3.4)$$

where $N_t(\mathbf{p}) = \sum_{\mathbf{d} \in \mathcal{D}} N_t(\mathbf{p}, \mathbf{d})$, and $N_t(\mathbf{p}, \mathbf{d})$ is the number of times up to day t that the algorithm simultaneously observes a daily target load profile $\mathbf{d}$ and posts a price $\mathbf{p}$. Furthermore, we define $\mathbf{N}_t = [N_t(\mathbf{p})]_{\mathbf{p} \in \mathcal{P}}$ as a vector including number of times each price is posted up to day t. We can consider the quantity in the exponent of (3.4) as a loss suffered by model $\boldsymbol{\theta}$ up to time t. Since the term in the exponent of (3.4) is equal to 0 when $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$, we can see that Thompson sampling samples $\boldsymbol{\theta}^\star$ and hence posts an optimal price with at least a constant probability at each day, i.e., $N_t(\mathbf{p}^{\star^\mathbf{d}}, \mathbf{d})$ grows linearly with t for all $\mathbf{d}$.

For each price, we define $S_\mathbf{p}(\mathbf{d}) := \{\boldsymbol{\theta} \in \Theta : \mathbf{p}_t = \mathbf{p}|\mathbf{d}_t = \mathbf{d}\}$ to be the set of parameters $\boldsymbol{\theta} \in \Theta$ whose optimal price (i.e., the solution of (3.1)) when observing a daily target load profile $\mathbf{d}$ is $\mathbf{p}$. Furthermore, define $S^{'}_\mathbf{p}(\mathbf{d}) := \{\boldsymbol{\theta} \in S_\mathbf{p}(\mathbf{d}) : D(\boldsymbol{\theta}^\star_{\mathbf{p}^{\star^\mathbf{d}}} \| \boldsymbol{\theta}_{\mathbf{p}^{\star^\mathbf{d}}}) = 0\}$ which is the set of models $\boldsymbol{\theta}$ that exactly match $\boldsymbol{\theta}^\star$ in marginal distribution of Y when the true model $\boldsymbol{\theta}^\star$ is selected and the optimal price $\mathbf{p}^{\star^\mathbf{d}}$ is posted, and $S^{''}_\mathbf{p}(\mathbf{d}) := S_\mathbf{p}(\mathbf{d}) \backslash S^{'}_\mathbf{p}(\mathbf{d})$.

For each model $\boldsymbol{\theta}$ in $S_{\mathbf{p}}^{''}(\mathbf{d})$, $\mathbf{p} \neq \mathbf{p}^{\star^{\mathbf{d}}}$, $D(\boldsymbol{\theta}^{\star}_{\mathbf{p}^{\star^{\mathbf{d}}}} \| \boldsymbol{\theta}_{\mathbf{p}^{\star^{\mathbf{d}}}}) > \varepsilon > 0$. As we have assumed that the probability of observing any target profile $\mathbf{d} \in \mathcal{D}$ is bounded away from zero, $N_t(\mathbf{p}^{\star^{\mathbf{d}}})$ grows linearly with t for all $\mathbf{d} \in \mathcal{D}$. Hence, any such model $\boldsymbol{\theta}$ is sampled with probability exponentially decaying in t in (3.4) and the regret from such $S_{\mathbf{p}}^{''}(\mathbf{d})$-sampling is negligible. We define the set of all such models as $\boldsymbol{\theta} \in \Theta'' = \cup_{\mathbf{d}\in\mathcal{D}} S_{\mathbf{p}}^{''}(\mathbf{d})$.

A model $\boldsymbol{\theta} \in S_{\mathbf{p}}^{'}(\mathbf{d})$ will only face loss whenever the algorithm posted a suboptimal price $\mathbf{p}$ for which $D(\boldsymbol{\theta}^{\star}_{\mathbf{p}} \| \boldsymbol{\theta}_{\mathbf{p}}) > 0$. For $\mathbf{d}$, a suboptimal price $\mathbf{p}_k^{\mathbf{d}} \neq \mathbf{p}^{\star^{\mathbf{d}}}$ may still be posted if any of the set of models in $S_{\mathbf{p}_k^{\mathbf{d}}}^{'}(\mathbf{d})$ may still be drawn with non-negligible probability. Hence, a price will be eliminated after the probability of drawing all $\boldsymbol{\theta} \in S_{\mathbf{p}_k^{\mathbf{d}}}^{'}(\mathbf{d})$ is negligible. For each $\mathbf{d}$, suboptimal prices are eliminated one after the other at times $t_k^{\mathbf{d}}, k = 1, \ldots, |\mathcal{P}| - 1$. The regret bounds we provide, which are adopted from [9] and generalized to take into account the effects of the daily profiles $\mathbf{d}$, characterize the total number of suboptimal prices posted as a function of T. The result holds under the assumptions that $|\mathcal{P}|, |\mathcal{Y}|, |\Theta| < \infty$ and the uniqueness of optimal price $\mathbf{p}^{\star^{\mathbf{d}}}$ for all $\mathbf{d} \in \mathcal{D}$. We refer the reader to [9] for a full discussion of when a suboptimal price $\mathbf{p}$ is considered statistically eliminated, which is used to write constraints (3.7)-(3.8) below.

Theorem 3.1. *For $\delta, \epsilon \in (0, 1)$, there exists $T^{\star} > 0$, such that for all $T > T^{\star}$, with probability at least $1 - \delta$,*

$$\sum_{\mathbf{d}\in\mathcal{D}} \sum_{\mathbf{p}\in\{\mathcal{P}\setminus\mathbf{p}^{\star^{\mathbf{d}}}\}} N_T(\mathbf{p}, \mathbf{d}) \leq B + C(logT).$$

Where $B \equiv B(\delta, \epsilon, \mathcal{P}, \mathcal{Y}, \Theta)$ is a problem-dependent constant which is not dependent on T, and:

$$C(log\ T) \equiv$$

$$\max \sum_{\mathbf{d}\in\mathcal{D}} \sum_{k=1}^{|\mathcal{P}|-1} N_{t_k^{\mathbf{d}}}(\mathbf{p}, \mathbf{d}) \tag{3.5}$$

$$s.t. \quad \forall \mathbf{d} \in \mathcal{D}, \forall j > 1, \forall 1 \leq k \leq |\mathcal{P}| - 1 : \tag{3.6}$$

$$\min_{\boldsymbol{\theta}\in\left\{S'_{\mathbf{p}_k^{\mathbf{d}}}(\mathbf{d})-\Theta''\right\}} \langle \mathbf{N}_{t_k^{\mathbf{d}}}, D_{\boldsymbol{\theta}} \rangle \geq \frac{1+\epsilon}{1-\epsilon} logT, \tag{3.7}$$

$$\min_{\boldsymbol{\theta}\in\left\{S'_{\mathbf{p}_k^{\mathbf{d}}}(\mathbf{d})-\Theta''\right\}} \langle \mathbf{N}_{t_k^{\mathbf{d}}} - e^{(j)}, D_{\boldsymbol{\theta}} \rangle < \frac{1+\epsilon}{1-\epsilon} logT, \tag{3.8}$$

where $e^{(j)}$ denotes the j-th unit vector in finite-dimensional Euclidean space. The last two constraints ensure that price $\mathbf{p}_k^{\mathbf{d}}$ is eliminated exactly at time $t_k^{\mathbf{d}}$ (no earlier and no later).

4. SIMULATION RESULTS

In our numerical experiment, we use the TS approach proposed in Algorithm 1 to quantify the potential cost reduction in a residential electric vehicle (EV) smart charging scenario under RTP. We assume our target profiles $\mathbf{d}$ belong to a set of 6 potential profiles representing general wind generation profiles at night time. We assume EVs can belong to one of 27 clusters, with the cluster parameters E_i, τ_1^i, and τ_2^i defined in (2.1). Vehicle charge requests are discretized in the simulation into seven 2-hour periods from 6 PM to 8 AM, periods of time particularly suited for residential EV charging. Taking a typical residential charging rate of 3.3kW, the period-wise charging rate ρ is specified as double this rate, resulting in ρ = 6.6 kWh/period. Charge requests range from 15 kWh to 45 kWh in 5-kWh increments. Time constraints τ_1^i, and τ_2^i were selected to produce charging periods of 10 hours, 12 hours, or 14 hours in length. The set of potential models adopted to represent the per cluster sensitivities of the users, i.e., $a_i(\mathbf{p})$, are selected as $a_i(\mathbf{p}) \sim \mathcal{N}(\frac{c_i}{\boldsymbol{\theta}^{\star T}\mathbf{p}}, \sigma^2)$, where c_i is a scalar specific to cluster i. The unknown parameter vector set Θ has 6 elements. Price signals $\mathbf{p}_t$ can be any vector of length seven with low ($\mathbf{p}(\tau) = 1$), medium ($\mathbf{p}(\tau) = 2$), and high ($\mathbf{p}(\tau) = 3$) price elements.

The cost function $g(.)$ is defined as a function of the deviation of the load profile $\mathbf{L}^{\star}(\mathbf{p_t})$ from the target profile $\mathbf{d}_t$. Upward deviations of the load from $\mathbf{d}_t$ were penalized at \$50 per MWh, whereas downward deviations of the load from $\mathbf{d}_t$ were penalized at \$30 per MWh (for each 2hr time period). The cumulative cost incurred by the aggregator over 100 days for a representative iteration is shown in Fig. 1. The cumulative cost incurred by Algorithm 1 is compared against a solution that knows $\boldsymbol{\theta}^{\star}$, i.e., optimizes posted prices with respect to the true model of the customers' price response.

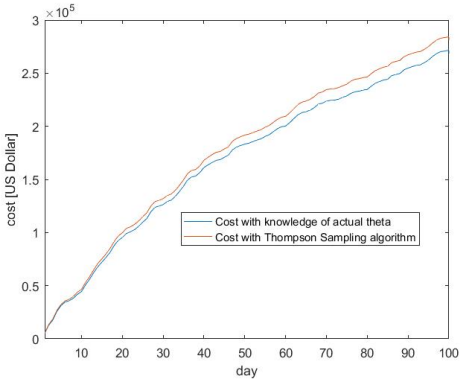

Fig. 1. Cumulative cost with knowledge of the true model $\boldsymbol{\theta}^{\star}$ versus that of Thompson sampling.

As the simulation proceeds, dynamic pricing decisions are assisted by the increasing certainty of the true model $\boldsymbol{\theta}^{\star}$ via Bayesian updating. In this way, the incurred cost with the Thompson sampling algorithm more closely matches the incurred cost with knowledge of the $\boldsymbol{\theta}^{\star}$.

5. REFERENCES

[1] A. Kavousian, R. Rajagopal, and M. Fischer, "Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior," *Energy*, vol. 55, pp. 184–194, 2013.

[2] R. Ganti, M. Sustik, Q. Tran, and B. Seaman, "Thompson sampling for dynamic pricing," *arxiv:1802.03050*, 2018.

[3] A. Krishnamurthy, Z. S. Wu, and V. Syrgkanis, "Semiparametric contextual bandits," *CoRR*, vol. abs/1803.04204, 2018.

[4] D. J. Foster, A. Agarwal, M. Dudík, H. Luo, and R. E. Schapire, "Practical contextual bandits with regression oracles," *CoRR*, vol. abs/1803.01088, 2018.

[5] T. Xu, Y. Yu, J. Turner, and A. Regan, "Thompson sampling in dynamic systems for contextual bandit problems," *CoRR*, vol. abs/1310.5008, 2013.

[6] D. Russo, B. V. Roy, A. Kazerouni, and I. Osband, "A tutorial on thompson sampling," *CoRR*, vol. abs/1707.02038, 2017.

[7] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," in *Mathematics of Operations Research*, vol. 39, no. 4, 2014, pp. 1221–1243.

[8] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Proceedings of the 25th Annual Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 23. Edinburgh, Scotland: PMLR, 2012, pp. 39.1–39.26.

[9] A. Gopalan, S. Mannor, and Y. Mansour, "Thompson sampling for complex online problems," in *International Conference on Machine Learning*, 2014, pp. 100–108.

[10] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 28, no. 3, Atlanta, Georgia, USA, 2013, pp. 127–135.

[11] Z. Xu, T. Deng, Z. Hu, Y. Song, and J. Wang, "Data-driven pricing strategy for demand-side resource aggregators," *IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 57–66, 2018.

[12] P. Li and B. Zhang, "Linear estimation of treatment effects in demand response: An experimental design approach," *CoRR*, vol. abs/1706.09835, 2017.

[13] P. Li, H. Wang, and B. Zhang, "A distributed online pricing strategy for demand response programs," *IEEE Transactions on Smart Grid*, pp. 1–1, 2017.

[14] Y. Li, Q. Hu, and N. Li, "Learning and selecting the right customers for reliability: A multi-armed bandit approach," 2018.

[15] K. Khezeli, W. Lin, and E. Bitar, "Learning to buy (and sell) demand response," *International Federation of Automatic Control (IFAC)*, vol. 50, no. 1, pp. 6761–6767, 2017.

[16] K. Khezeli and E. Bitar, "Data-driven pricing of demand response," pp. 224–229, Nov 2016.

[17] L. Jia, L. Tong, and Q. Zhao, "An online learning approach to dynamic pricing for demand response," *arXiv:1404.1325*, 2014.

[18] G. S. Ledva, L. Balzano, and J. L. Mathieu, "Real-time energy disaggregation of a distribution feeder's demand using online learning," *IEEE Transactions on Power Systems*, 2018.

[19] D. Kalathil and R. Rajagopal, "Online learning for demand response," pp. 218–222, Sept 2015.

[20] M. Alizadeh, A. Scaglione, A. Applebaum, G. Kesidis, and K. Levitt, "Reduced-order load models for large populations of flexible appliances," *IEEE Transactions on Power Systems*, vol. 30, no. 4, pp. 1758–1774, 2015.