Retail pricing for stochastic demand with unknown parameters: an online machine learning approach

Liyan Jia, Qing Zhao and Lang Tong

***Abstract*— The problem of dynamically pricing of electricity by a retailer for customers in a demand response program is considered. It is assumed that the retailer obtains electricity in a two-settlement wholesale market consisting of a day ahead market and a real-time market. Under a day ahead dynamic pricing mechanism, the retailer aims to learn the aggregated demand function of its customers while maximizing its retail profit. A piecewise linear stochastic approximation algorithm is proposed. It is shown that the accumulative regret of the proposed algorithm grows with the learning horizon T at the order of $O(\log T)$. It is also shown that the achieved growth rate cannot be reduced by any piecewise linear policy.**

***Index Terms*— Demand response; electricity retail pricing; online learning; stochastic approximation; optimal stochastic thermal control.**

I. INTRODUCTION

We consider the problem of pricing of electricity by a retailer for customers who participate in a demand response program but whose demand functions are unknown. We assume that the retailer obtains electricity from a two-settlement electricity market where the retailer receives a financially binding day ahead schedule in terms of the day ahead cleared price and quantity. In real time, the retailer serves its customers by purchasing electricity in the wholesale market, and the amount of electricity deviated from the day ahead schedule is settled according to the real-time wholesale price.

We assume that the retailer can influence the demand of its customers through some form of real-time pricing. If the retailer knows how its customers response to price through their individual demand functions, the retailer can choose its price to optimize a its objective, *e.g.,* the social welfare or its own profit. Obtaining the demand functions of its customers, however, is nontrivial because a customer is likely to consider such information private; neither the willingness of sharing nor the truthfulness of the shared information can be assumed.

We formulate this problem as one of online learning problems where the retailer learns the behavior of its customers by observing the response of its customers to carefully designed prices. The basic principle of online learning is to achieve a tradeoff between “exploration” and “exploitation”; the former represents the need of using sufficiently rich pricing signals to achieve accuracy of learning, whereas the latter represents the need of capturing as much reward as possible based on what has been learned.

In the classical online learning theory, the performance of a learning algorithm can be measured by the notion of accumulative regret. For the pricing problem at hand, the regret is defined by the difference between the “oracle profit” associated with the actual demand function and the profit achieved by the online learning algorithm. While the accumulative regret R_T grows with the learning horizon T, the rate of regret R_T/T should diminish, which implies that, for the infinite horizon problem, the profit achieved per unit time without knowing the demand function matches that when the demand function is known. However, because the consumption of electricity depends on environmental and behaviorial patterns, it is more relevant to consider the finite horizon problem. To this end, the appropriate measure is the rate of convergence, $i.e.,$ the rate at which R_T/T decays to zero, or equivalently, how slowly R_T grows with T. One would prefer an algorithm whose accumulative regret grows at the order of $O(\log(T))$ rather than $O(T^\alpha)$.

A. Summary of results

This paper presents an application of on-line learning theory tailored for the problem of pricing of electricity for distribution customers who participate in a demand response program. We focus on thermal dynamic loads for which electricity is consumed to maintain temperature near preferred settings. The retailer may have many such customers. We assume that the retailer does not know the desired temperature set-points, nor the parameters that characterize the thermal dynamics of their environments.

We assume that the retailer employs a widely used real-time pricing scheme, referred to as day ahead dynamic price (DADP), under which the retailer posts the hourly price of electricity 24 hour ahead of time. First proposed by Borenstein, Jaske, and Rosenfield [1], this pricing scheme has been implemented in practice [1], [2].

A key advantage of DADP is that a customer has the short-term (24 hours) price certainty with which it can optimize its consumption. For thermal dynamic load, it is shown that the aggregated thermal load is an affine function of the 24 hour pricing vector π. This result, first shown in [3], characterizes the customer behavior except that parameters of the affine function is unknown to the retailer.

This work is supported by the National Science Foundation under Grant CNS-1135844 and CNS-1248079.

Liyan Jia and Lang Tong are with School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA, 14850. Email: {lj92,lt35}@cornell.edu.

Qing Zhao is with School of Electrical and Computer Engineering, University of California, Davis, CA, USA, 95616. Email: qzhao@ucdavis.edu.

We show in this paper that, in a two-settlement market, maximizing retail profit can be achieved by minimizing the 2-norm deviation of real-time demand from the day-ahead schedule. As a result, the problem becomes one of tracking with unknown system parameters.

Assuming that demand level determined by the day-ahead market is discrete, we propose a piecewise linear stochastic approximation (PWLSA) policy, as a generalization of an approach first proposed by Lai and Robbins in [4]. Specifically, the policy maintains adaptively a dictionary $\{(D_i, \mu_t^i)\}$ where D_i is the day ahead demand level and μ_t^i is a linear stochastic approximation pricing policy associated with D_i at time t. Given the day-ahead dispatch $d_t^{\rm DA} = D_i$, the PWLSA pricing policy μ_t^i generates the real-time retail price π_t.

We show that the accumulative regret of the PWLSA pricing policy grows with the learning horizon T at the order of $\Theta(\log T)$. We show further that any piecewise (time varying) linear policy cannot have the accumulative regret grow at $o(logT)$.

B. Related work

The problem of dynamic pricing for demand response assuming known demand function has been extensively studied. See [1], [5], [2] for discussions of the pricing scheme considered in here and [6], [7], [8], [9] for more general settings. These results assume implicitly that the demand function is known. A precursor of the work presented here is [3] where a parametric form of demand function is obtained. In [10], the tradeoff between retail profit and consumer surplus is characterized under a Stackelberg formulation under the assumption that the demand functions of customers are known.

Online learning of unknown demand functions has been studied extensively in multiple communities. This problem can be formulated as a multi-armed bandit (MAB) problem by treating each possible price as an arm. When the price can only take finite possible values, the problem becomes the classic MAB for which Lai and Robbins showed that the optimal regret growth rate is $\Theta(\log T)$ when the arms generate independent reward [11]. When the price takes value from an uncountable set, the dynamic pricing problem is an example of the so-called continuum-armed bandit introduced by Agrawal in [12] where the arms form a compact subset of $\mathcal{R}$. An online learning policy with regret order of $O(T^{3/4})$ was proposed in [12] for any reward function satisfying Lipschitz continuity. Further development on the continuum-armed bandit under various assumptions of the unknown reward function can be found in [13], [14], [15]. The reason that PWLSA proposed in this paper achieves a much better regret order ($O(\log T)$) than in the case of a general continuum-armed bandit is due to the specific linearly parameterized demand which leads to a specific quadratic cost/reward function. A similar message can be found in [16], [17], [18] where different regret orders were shown to be achievable under different classes of demand models for dynamic pricing.

The problem considered in this paper deals with linearly parameterized demand function, thanks to the closed-form characterization of the optimized demand function for thermal dynamic load. The learning approach proposed in this paper is rooted from the stochastic approximation problem originally formulated by Lai and Robbins [19], [4] where the authors considered a form of optimal control problem when the model contains unknown parameters and the cost of control is explicitly modeled. For scaler models, the authors of [19], [4] showed that the cumulative regret (if translated from our definition) of a simple linear stochastic approximation scheme grows at the rate of $O(\log T)$. However, it is not clear whether such growth rate is the lowest possible. Our result provides a generalization to the vector case with a lower bound for a general class of piecewise linear policies of which linear stochastic approximation is a special case.

Also related is the work of Bertsimas and Perakis [20] who tackled the problem as a dynamic program with incomplete state information. The authors showed in numerical simulations that considerable gain can be realized over the myopic policy where the price in the next stage is based on the least squares estimate of the model parameter. When the parameters are assumed to be random, Lobo and Boyd considered same problem here under a Bayesian setting [21]. The authors introduced a randomization policy via a dithering mechanism.

Machine learning techniques have been applied to pricing problems in electricity markets, although there seems to be limited literature on discovering real-time price with unknown known demand functions at retail level. While such problems can be viewed as part of the general learning problem discussed above, the nature of electricity market and electricity demand impose special constraints. A related learning problem of bidding strategy of a retailer in the whole sale market when the supply functions of the generators are unknown has been studied. See [22], [23], [24] where Q-learning techniques have been applied.

II. The two-settlement wholesale market

Most US deregulated wholesale electricity markets adopt a two-settlement system consisting of a day-ahead market and a real-time market. We describe in this section the participation of a retailer in the wholesale market and argue that, if the retailer is to influence the consumption of its customers via retail pricing in real-time, the profit maximizing strategy is to choose the retail price to minimize the 2-norm deviation between the day ahead scheduled demand and the real-time counter part. This is not surprising except perhaps that the 2-norm measure of deviation is the appropriate metric.

A. The day-ahead wholesale market

In the day-ahead market, a retailer (or a Load Serving Entity (LSE)) submits a utility curve $u(d)$ that represents the benefit of getting served with d units electricity in the second day. An electricity generator, on the other hand, submits a cost curve $c(p)$ that represents the cost of serving p units electricity in the next day. Because the day-ahead market is defined at the hourly time scale, the demand schedule d and generation schedules p are 24 dimension vectors.

The independent system operator (ISO) aims to maximize the social welfare by solving an optimal power flow (OPF) problem. In its simplest form without complications of a capacity constrained transmission network and multiple participating agents, the OPF problem is of the following form

$$\begin{aligned} \max \quad & u(d) - c(p) \\ \text{s.t.} \quad & d = p \end{aligned}$$

Let the solution of the above optimization $d^{\text{DA}} = p^{\text{DA}}$ be the cleared day ahead dispatch. The day-head cleared price is the marginal cost of generating p^{DA}, calculated as $\lambda^{\text{DA}} = \frac{\partial c}{\partial p}(p^{\text{DA}})$. Note that the clearing of the day-ahead market is financially binding, in the sense that regardless of the actual consumption in the real time, the day-ahead payment from retailer to the system operator is $(\lambda^{\text{DA}})^T d^{\text{DA}}$ and the payment from system operator to generator is $(\lambda^{\text{DA}})^T p^{\text{DA}}$.

The retail surplus of a retailer can be illustrated in the Price-Quantity plane. As shown in Fig. 1, the day-ahead equilibrium is $(d^{\text{DA}}, \lambda^{\text{DA}})$ and Area I represents the day-ahead retail surplus [25].

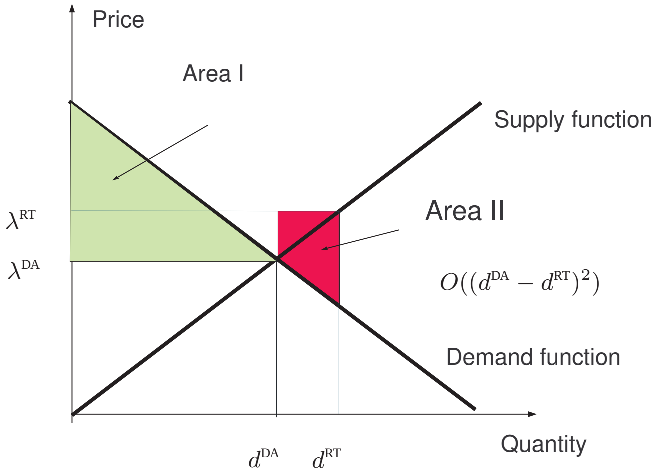

Fig. 1: The day ahead and real-time market equilibria

B. *The real-time wholesale market*

In the real time wholesale market, the actual consumption, d^{RT}, may be different from the day ahead dispatch, which affects the wholesale price in the real-time market. In particular, the real-time price may be calculated according to $\lambda^{\text{RT}} = \frac{\partial c}{\partial p}(d^{\text{RT}})$. Here, for simplicity, we assume that the cost function of the generator in real-time is the same as that in the day ahead market. In practice, however, the cost functions are not necessarily the same. The intuitions behind our arguments remain.

In the real-time market, payments from the retailer to system operator and from system operator to generator are both $(\lambda^{\text{RT}})^T(d^{\text{RT}} - d^{\text{DA}})$ if this quantity is positive. Otherwise, the compensations from system operator to retailer and from generator to system operator are both $(\lambda^{\text{RT}})^T(d^{\text{DA}} - d^{\text{RT}})$.

In the real-time market, if the real-time consumption $d^{\text{RT}} = d^{\text{DA}}$, there is no real-time payment. The total retail surplus is Area I in Fig. 1. However, if the real-time consumption d^{RT} is different from d^{DA}, as shown in Fig. 1, the real-time price, λ_{RT}, is determined by the supply function. Hence, the total retail surplus is,

$$u(d^{\text{RT}}) - d^{\text{DA}}\lambda^{\text{DA}} - (d^{\text{RT}} - d^{\text{DA}})\lambda^{\text{RT}},$$

which can be represented by the area difference between the Area I and Area II. Therefore, Area II is the retail surplus loss, and the loss grows in the order of $(d^{\text{RT}} - d^{\text{DA}})^2$—the deviation between the day-ahead scheduled consumption and the actual real-time consumption.

For the general vector case, the result is formally expressed in the following theorem.

Theorem 1: Under the two-settlement market system, if the generation cost function $c(\cdot)$ has a quadratic form, $c(d) = d^T K d + h^T d + c(0)$, where K is p.s.d., then the retailer's surplus loss is approximately measured by $2(d^{\text{RT}} - d^{\text{DA}})^T K (d^{\text{RT}} - d^{\text{DA}})$.

Proof: see Appendix.

In practice, K is usually diagonal. In the later discussion, without loss of generality, we assume $K = I$. Then, the objective of the retailer is to minimizing squared deviation from the real-time demand to day-ahead dispatch. The expected surplus loss for the t-th day can be defined as

$$L_t \overset{\Delta}{=} \mathbb{E}[||d_t^{\text{RT}} - d_t^{\text{DA}}||_2^2].$$

where d_t^{DA} and d_t^{RT} are the day-ahead dispatch and real-time demand for day t.

III. Dynamic Pricing in the Retail Market

In this section, we describe interactions between a retailer and its customers. It is assumed that the retailer has received the day-schedule for the committed day ahead quantity, denoted here as d_t^{DA} for day t. Without knowing the demand functions of its customers, the retailer aims to optimize its surplus by adaptively choosing the real-time price for its consumers.

A. *Day-ahead dynamic pricing*

In this paper, we focus on a particular class of pricing mechanisms, referred to as day-ahead dynamic pricing (DADP), to control the demand response. The principle of DADP is that the retailer posts day-ahead hourly prices, and these prices will be fixed at the day of consumption. First considered in [1], DADP has been in place for large retail customers for years. The advantage of DADP for a consumer is that the consumer has the price certainty one day ahead of time so that he can plan accordingly based on the posted prices and his desired quality of service.

The sequence of the retail market operations under the DADP pricing is as follows:

- The retailer offers the consumer day ahead hourly retail price π and keeps it fixed for the next day.
- In real-time, a consumer optimizes its consumption based on π.

- The payment from a consumer to the retailer is settled as the product of π and real-time consumption.
- The retailer meets aggregated demand by purchasing electricity at the wholesale market and pay the real-time wholesale price for the deviation from the day ahead amount.

B. Optimal demand respose

We consider in this section the optimization of the demand response to DADP. In practice, thermal loads (HVACs units[*]) represent a significant part of price responsive demand. Empirical study [26] has shown that the dynamic equation that governs the HVAC temperature evolution is given by

$$x_i = x_{i-1} + \alpha(a_i - x_{i-1}) - \beta u_i + \xi_i, \tag{1}$$

where $x = (x_1, x_2, ..., x_{24})$ is the vector of average outdoor temperature in each hour, $a = (a_1, a_2, ..., a_{24})$ is the vector of average outdoor temperature in each hour, $u = (u_1, ..., u_{24})$ the vector of control variable representing the total amount of electricity drawn by the HVAC unit during each hour and $\xi = (\xi_1, \xi_2, ..., \xi_{24})$ the process noise. System parameters α $(0 < \alpha < 1)$ and β model the insolation of the building and efficiency of the HVAC unit. Note that the above equation applies to both heating and cooling scenarios but not simultaneously. We focus herein the cooling scenario $(\beta > 0)$ and the results apply to heating $(\beta < 0)$ as well.

Using a linear combination of total cost and squared deviation of indoor temperature from desired temperature as the objective function, the optimized demand response from the customer can be formulated as the following stochastic optimization problem,

$$\begin{array}{rl} \min & \mathbb{E}\left\{\sum_{i=1}^{24}[-\mu(x_i - t_i)^2] - \pi^{\mathrm{T}} u\right\} \\ \text{s.t.} & x_i = x_{i-1} + \alpha(a_i - x_{i-1}) - \beta u_i + \xi_i, \\ & y_i = (x_i, a_i) + \nu_i. \end{array}$$

where $y = (y_1, y_2, ..., y_{24})$ is the observation vector, $\nu = (\nu_1, \nu_2, ..., \nu_{24})$ the observation noise vector.

The solution of the above stochastic optimization can be obtained in closed form via a direct backward induction. More significantly, it is shown in [10] that the total demand is a linear function of the retail price.

Theorem 2 ([10]): Assume that the process noise ξ and ν are Gaussian distributed with zero mean. For fixed retail price π, the optimal aggregated residential demand response has the following matrix form and properties

$$d^{\text{RT}} = b - A\pi + w, \tag{2}$$

where the factor matrix A is positive definite, b and A are deterministic, depending only on the dynamic system parameters, and w is a random vector with zero mean.

IV. Dynamic pricing via online machine learning

As described in Section II-B, the retailer's objective is to minimize the surplus loss, defined as the squared deviation of real-time electricity consumption, d^{RT}, from the day-ahead optimal dispatch, d^{DA}.

If the parameters, A and b, are known to the retailer in the demand model (2), at day t, given the day-ahead dispatch d_t^{DA}, the optimal retail price should be designed as

$$\begin{array}{rl} \pi_t^* & = \arg\min_{\pi_t} \mathbb{E}[||d_t^{\text{RT}} - d_t^{\text{DA}}||_2^2] \\ & = \arg\min_{\pi_t} \mathbb{E}[||b - A\pi_t + w - d_t^{\text{DA}}||_2^2] \\ & = A^{-1}(b - d_t^{\text{DA}}). \end{array} \tag{3}$$

The minimum of the expected surplus loss is

$$\mathbb{E}[||b - A\pi_t^* + w - d_t^{\text{DA}}||_2^2] = \Sigma_w,$$

where Σ_w is the covariance matrix of demand model noise w in (2). Notice that the minimized surplus loss is independent of the day-ahead dispatch d_t^{DA}.

However, it is nontrivial for the retailer to obtain the parameters of the demand functions of its customers because a customer is likely to consider such information private. At day t, the only information available to the retailer is the record of previous electricity consumption up to $t-1$ and day-ahead dispatch up to t. Hence, the retail pricing policy μ is defined as

$$\pi_t^{\mu} = \mu_t(d_1^{\text{RT}}, ..., d_{t-1}^{\text{RT}}, d_1^{\text{DA}}, ..., d_{t-1}^{\text{DA}}, d_t^{\text{DA}}) \tag{4}$$

where d_i^{DA}, and d_i^{RT} are the day-ahead dispatch and real-time electricity consumption for day i.

A. Piecewise Linear Stochastic Approximation policy

In [4], Lai and Robbin show that if π_t and d_t^{RT} are both scalars and $d_t^{\text{DA}} = d^{\text{DA}}$ is constant for all t, the stochastic approximation policy,

$$\pi_t = \bar{\pi}_{t-1} + \gamma(\bar{d}_{t-1}^{\text{RT}} - d^{\text{DA}}), \tag{5}$$

achieves $\pi_t \rightarrow \pi^*$ a.s. and $O(\log(T))$ aggregated regret (defined in Section IV-B), where π^* is the optimal price in (3) with day-ahead dispatch d^{DA}, $\bar{\pi}_{t-1}$ is the average of $\pi_1, ..., \pi_{t-1}$ and $\bar{d}_{t-1}^{\text{RT}}$ is the average of $d_1^{\text{RT}}, ..., d_{t-1}^{\text{RT}}$.

In this paper, we propose a policy named *Piecewise Linear Stochastic Approximation (PWLSA)*, extending the stochastic approximation policy to multi-dimensional case with countable many day-ahead dispatch levels.

The basic idea is that we record all the day-ahead dispatch levels up to day t in set $\mathcal{D}_t$. For day $t+1$, if $d_{t+1}^{\text{DA}} \in \mathcal{D}_t$, we let $\mathcal{D}_{t+1} = \mathcal{D}_t$. Otherwise, $\mathcal{D}_{t+1} = \mathcal{D}_t \bigcup \{d_t^{\text{DA}}\}$. For each day-ahead dispatch level in $\mathcal{D} = \bigcup \mathcal{D}_t$, we keep a separate stochastic approximation to calculate the retail price.

Therefore, for different d_t^{DA}, we have a different linear function to calculate the next retail price. The policy is piecewise linear. Formally, the PWLSA policy, μ^{PWLSA}, is defined as,

Definition 1 (PWLSA): Assume for all $t \in \mathbb{N}^+$, $d_t^{\text{DA}} \in \mathcal{D}$ and $\mathcal{D}$ is countable.

[*] Heating, ventilation, and air conditioning units

- If $d_t^* = D_i \in \mathcal{D}_t$, $\mathcal{D}_{t+1} = \mathcal{D}_t$,

$$\pi_t^{\text{PWLSA}} = \frac{1}{|\mathcal{C}_t^i|}\left(\sum_{k\in\mathcal{C}_i} \pi_k^{\text{PWLSA}} + \gamma(d_k^{\text{PWLSA}} - d_t^{\text{DA}})\right)$$

  where $\mathcal{C}_t^i = \{k \in \mathbb{N}^+ : k \leq t-1, d_k^{\text{DA}} = D_i\}$
- Otherwise, $\mathcal{D}_{t+1} = \mathcal{D}_t \bigcup \{d_t^{\text{DA}}\}$

B. Regret analysis of PWLSA

For each policy μ as in (4), the regret at day t, R_t^μ, is defined as the difference of retail's expected surplus loss between using π_t^μ and π_t^*. Therefore, according to (3),

$$\begin{aligned} R_t^\mu \quad &\triangleq \mathbb{E}[||b - A\pi_t^\mu + w_t - d_t^{\text{DA}}||_2^2 - \Sigma_w \\ &= \mathbb{E}[||b - A\pi_t^\mu - d_t^{\text{DA}}||_2^2] \end{aligned} \tag{6}$$

Now we want to use the regret to evaluate the performance of PWLSA. First, we show that PWLSA can achieve $\log(T)$ accumulated regret under certain conditions.

Theorem 3: Assume that day-ahead dispatch d_t^{DA}'s are from a finite set, $i.e.,\ |\mathcal{D}| < \infty$. If $\gamma \geq \frac{1}{2\lambda_{\min}(A)}$,

$$\sum_{t=1}^{T} R_t^{\mu^{\text{PWLSA}}} \sim O(\log(T))$$

Proof: see Appendix.

Now we focus on those piecewise linear pricing policies, whose changing points are independent of d_i^{RT}, $i = 1, 2, ...$. Formally, the set of such polices, $\mathcal{P}$, is defined as,

$$\mathcal{P} = \{\mu : \pi_t = \mu_t(d_1^{\text{RT}}, ...d_{t-1}^{\text{RT}}; d_1^{\text{DA}} ..., d_t^{\text{DA}})\} \tag{7}$$

where $\mu_t(\cdot; d_1^{\text{DA}}, ..., d_t^{\text{DA}})$ is a linear function of $d_1^{\text{RT}}, ...d_{t-1}^{\text{RT}}$.

Clearly, our proposed policy $\mu^{\text{PWLSA}} \in \mathcal{P}$. The following theorem shows that $\log(T)$ accumulated regret is the best rate that can be achieved by any policy in $\mathcal{P}$.

Theorem 4: For any policy $\mu \in \mathcal{P}$,

$$\sum_{t=1}^{T} R_k^\mu \geq C^\mu \log(T)$$

Proof: see Appendix.

V. CONCLUSION AND FUTURE WORK

We present in this paper an online learning approach to the dynamic pricing of electricity of a retailer whose customers have price responsive dynamic loads with unknown demand functions. We exploit the linear form of the aggregated demand function, and cast the problem of online learning as one of tracking day-ahead wholesale prices via a stochastic optimization. This approach leads to a simple learning algorithm with the growth rate of accumulative regret at the order of $O(\log T)$, which is the best rate achievable for a class of piecewise linear policies.

REFERENCES

[1] S. Borenstein, M. Jaske, and A. Rosenfeld, "Dynamic Pricing, Advanced Metering, and Demand Response in Electricity Markets," *Recent Work, Center for the Study of Energy Markets, University of California Energy Institute, UC Berkeley*, 2002.

[2] N. Hopper, C. Goldman, and B. Neenan, "Demand Response from Day-Ahead Hourly Pricing for Large Customers," *Electricity Journal*, no. 02, pp. 52–63, 2006.

[3] L. Jia and L. Tong, "Optimal pricing for residential demand response: A stochastic optimization approach," in *2012 Allerton Conference on Communication, Control and Computing*, Oct. 2012.

[4] T. L. Lai and H. Robbins, "Iterated Least Squares in Multiperiod Control," *Advanced and Applied Mathematics*, vol. 3, pp. 50–73, 1982.

[5] S. Borenstein, "The long run efficiency of real-time electricity pricing," *The Energy Journal*, vol. 26, no. 3, 2005.

[6] M. Carrion, A. Conejo, and J. Arroyo, "Forward Contracting and Selling Price Determination for a Retailer," *IEEE Transactions on Power Systems*, vol. 32, no. 4, November 2007.

[7] A. Conejo, R. Garcia-Bertrand, M. Carrion, A. Caballero, and A. Andres, "Optimal Involvement in Futures Markets of a Power Producer," *IEEE Transactions on Power Systems*, vol. 23, no. 2, May 2008.

[8] P. Yang, G. Tang, and A. Nehorai, "A Game-Theoretic Approach for Optimal Time-of-Usa Electricity Pricing," *IEEE Transactions on Power Systems*, 2012.

[9] J. Y. Joo and M. Ilic, "Multi-Layered Optimization Of Demand Resources Using Lagrange Dual Decomposition," *IEEE Transactions on Smart Grid*, 2013.

[10] L. Jia and L. Tong, "Day ahead dynamic pricing for demand response in dynamic environments," in *52nd IEEE Conference on Decision and Control*, Dec. 2013.

[11] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advanced and Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[12] R. Agrawal, "The Continuum-Armed Bandit Problem," *SIAM J. Control and Optimization*, vol. 33, no. 6, pp. 1926–1951, 1995.

[13] R. Kleinberg, "Nearly Tight Bounds for the Continuum-Armed Bandit Problem," *Advances in Neural Information Processing Systems*, pp. 697–740, 2004.

[14] P. Auer, R. Ortner, and C. Szepesvari, "Improved Rates for the Stochastic Continuum-Armed Bandit Problem," *Lecture Notes in Computer Science*, vol. 4539, pp. 454–468, 2007.

[15] E. W. Cope, "Regret and Convergence Bounds for a Class of Continuum-Armed Bandit Problems," *IEEE Transactions on Automatic Control*, vol. 54, no. 6, Jan. 2009.

[16] R. Kleinberg and T. Leighton, "The value of knowing a demand curve: bounds on regret for online posted-price auctions," in *Proc. 44th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2003.

[17] P. Rusmevichientong and J. N. Tsitsiklis, "Linearly Parameterized Bandits," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 395–411, 2010.

[18] J. Broder and P. Rusmevichientong, "Dynamic Pricing under a General Parametric Choice Model," *Operations Research*, vol. 60, no. 4, pp. 965–980, 2012.

[19] T. L. Lai and H. Robbins, "Adaptive design and stochastic approximation," *The Annals of Statistics*, vol. 7, no. 6, pp. 1196–1221, 1979.

[20] D. Bertsimas and G. Perakis, "Dynamic Pricing: A Learning approach," *Mathematical and Computational Models for Congestion Charging*, pp. 45–80, 2006.

[21] M. Lobo and S. Boyd, "Pricing and learning with uncertain demand," in *INFORMS Revenue Management Conference*, Columbia University, 2003.

[22] A. Rahimi-Kian, B. Sadeghi, and R. J. Thomas, "Q-learning based supplier-agents for electricity markets," in *IEEE Power Engineering Society General Meeting*, 2005.

[23] Z. Qiu, E. Peeters, and G. Deconinck, "Comparison of Two Learning Algorithms in Modelling the Generator's Learning Abilities," in *15th International Conference on Intelligent System Applications to Power Systems*, 2009.

[24] T. Pinto, Z. Vale, F. Rodrigues, and I. Praca, "Cost dependent strategy for electricity markets bidding based on adaptive reinforcement learning," in *the 16th International Conference onIntelligent System Application to Power Systems*, 2011.

[25] A. Mas-Colell and M. D. Whinston, *Microeconomics Theory*. Oxford University Press, 1995.

[26] D. Bargiotas and J. Birddwell, "Residential air conditioner dynamic model for direct load control," *IEEE Transactions on Power Delivery*, vol. 3, no. 4, pp. 2119–2126, Oct. 1988.

Appendix

Proof of Theorem 1

Given the day-ahead demand d^{DA} and real-time demand d^{RT}, the demand side surplus is

$$u(d^{\text{RT}}) - (\lambda^{\text{DA}})^T d^{\text{DA}} - (\lambda^{\text{RT}})^T (d - d^{\text{DA}}).$$

Therefore, the surplus loss due to deviation of d^{RT} from d^{DA} is

$$\begin{array}{rcl} \text{Loss} & = & [u(d^{\text{DA}}) - (\lambda^{\text{DA}})^T d^{\text{DA}}] \\ & & -[u(d^{\text{RT}}) - (\lambda^{\text{DA}})^T d^{\text{DA}} - (\lambda^{\text{RT}})^T (d^{\text{RT}} - d^{\text{DA}})] \\ & = & u(d^{\text{DA}}) - u(d^{\text{RT}}) - (\lambda^{\text{RT}})^T (d^{\text{DA}} - d^{\text{RT}}). \end{array}$$

Consider the first order approximation

$$u(d^{\text{DA}}) - u(d^{\text{RT}}) \approx [\frac{\partial u}{\partial d}(d^{\text{DA}})]^T (d^{\text{RT}} - d^{\text{DA}}).$$

At the optimal day ahead dispatch, $d^{\text{DA}} = p^{\text{DA}}$ and

$$\frac{\partial u}{\partial d}(d^{\text{DA}}) = \frac{\partial c}{\partial p}(d^{\text{DA}}).$$

Hence,

$$\begin{array}{rl} \text{Loss} \approx & (\frac{\partial c}{\partial p}(d^{\text{RT}}) - \frac{\partial c}{\partial p}(d^{\text{DA}}))^T (d^{\text{RT}} - d^{\text{DA}}) \\ = & (d^{\text{RT}} - d^{\text{DA}})^T K (d^{\text{RT}} - d^{\text{DA}}). \end{array}$$

■

Proof of Theorem 3

First, we consider the case when there is a single day-ahead dispatch level d^{DA}, and $\pi^* = A^{-1}(b - d^{\text{DA}})$.

From the policy, we have

$$\bar{\pi}_{i+1} = \frac{1}{i+1}[\pi_{i+1} + i\bar{\pi}_i] = (1 - \frac{\gamma A}{i+1})\bar{\pi}_i + \frac{\gamma}{i+1}[A\pi^* + \bar{\omega}_i].$$

Therefore,

$$\begin{array}{rcl} \bar{\pi}_{i+1} - \bar{\pi}^* & = & [\Pi_{k=1}^{i}(1 - \frac{\gamma A}{k+1})](\pi_1 - \pi^*) \\ & & + \sum_{j=1}^{i} \frac{\gamma}{j+1}[\Pi_{k=j+1}^{i}(1 - \frac{\gamma A}{k+1})]\bar{\omega}_j. \end{array}$$

Therefore,

$$\begin{array}{l} \pi_{n+1} - \pi_{n+1}^* \\ = (n+1)(\bar{\pi}_{n+1} - \pi^*) - n(\bar{\pi}_n - \pi^*) \\ = (1 - \gamma A)[\Pi_{i=1}^{n-1}(1 - \frac{\gamma A}{i+1})](\pi_1 - \pi^*) \\ \quad + \sum_{k=1}^{n}\{\frac{\gamma}{n} + \sum_{j=k}^{n-1}[\Pi_{i=j+1}^{n-1}(1 - \frac{\gamma A}{i+1})]\frac{(1-\gamma A)\gamma}{j(j+1)}\}\omega_k. \end{array} \tag{8}$$

For the first term in (8),

$$\begin{array}{l} ||(1 - \gamma A)[\Pi_{i=1}^{n-1}(1 - \frac{\gamma A}{i+1})]||_2^2 \\ \leq ||(1 - \gamma A)||_2^2 \Pi_{i=1}^{n-1}||(1 - \frac{\gamma A}{i+1})||_2^2. \end{array}$$

Since

$$(1 - \frac{\gamma A}{i+1})^T(1 - \frac{\gamma A}{i+1}) = 1 - \frac{2\gamma A}{i+1} + \frac{\gamma^2 A^2}{(i+1)^2},$$

denoting λ_m as the minimum eigenvalue of A, we have,

$$||(1 - \frac{\gamma A}{i+1})||_2^2 \leq 1 - \frac{2\gamma\lambda_m}{i+1} + \frac{\gamma^2}{(i+1)^2}||A||_2^2.$$

Let $C_1 \overset{\Delta}{=} ||1 - \gamma A||_2^2$. Then, since $\gamma\lambda_m > \frac{1}{2}$

$$\begin{array}{l} ||(1 - \gamma A)[\Pi_{i=1}^{n-1}(1 - \frac{\gamma A}{i+1})]||_2^2 \\ \leq C_1 \Pi_{i=1}^{n-1}(1 - \frac{2\gamma\lambda_m}{i+1} + \frac{\gamma^2}{(i+1)^2}||A||_2^2) \\ \leq C_1 \text{exp}\{\sum_{i=1}^{n} - \frac{2\gamma\lambda_m}{i+1} + \frac{\gamma^2}{(i+1)^2}||A||_2^2\} \\ \leq C_1 \text{exp}\{-\log(n+1) + \gamma^2||A||_2^2\} \\ = C_2 \frac{1}{n+1}, \end{array}$$

where $C_2 = C_1 \text{exp}\{\gamma^2||A||_2^2\}$ doesn't depend on n.

For the second term in (8),

$$\begin{array}{l} ||[\Pi_{i=j+1}^{n-1}(1 - \frac{\gamma A}{i+1})](1 - \gamma A)||_2^2 \\ \leq C_1 \text{exp}\{\sum_{i=j+1}^{n} - \frac{2\gamma\lambda_m}{i+1} + \frac{\gamma^2}{(i+1)^2}||A||_2^2\} \\ \leq C_2(\frac{j+1}{n+1})^{2\gamma\lambda_m} \end{array}$$

Then,

$$\begin{array}{l} ||\frac{\gamma}{n} + \sum_{j=k}^{n-1}[\Pi_{i=j+1}^{n-1}(1 - \frac{\gamma A}{i+1})]\frac{(1-\gamma A)\gamma}{j(j+1)}||_2^2 \\ \leq \{\frac{\gamma}{n} + \sum_{j=k}^{n-1}||[\Pi_{i=j+1}^{n-1}(1 - \frac{\gamma A}{i+1})](1 - \gamma A)||_2 \frac{\gamma}{j(j+1)}\}^2 \\ \leq \{\frac{\gamma}{n} + \sum_{j=k}^{n-1}\sqrt{C_2}(\frac{j+1}{n+1})^{\gamma\lambda_m}\frac{\gamma}{j(j+1)}\}^2 \\ \leq \{\frac{\gamma}{n} + \sum_{j=k}^{n-1}\gamma\sqrt{C_2}(\frac{1}{n})^{\gamma\lambda_m}(\frac{1}{j})^{2-\gamma\lambda_m}\}^2 \\ \leq \{\frac{\gamma}{n} + \gamma\sqrt{C_2}(\frac{1}{n})^{\gamma\lambda_m}(\frac{1}{k})^{1-\gamma\lambda_m}\}^2 \\ \leq 2\frac{\gamma^2}{n^2} + 2\gamma C_2(\frac{1}{n})(\frac{1}{n})^{2\gamma\lambda_m - 1}(\frac{1}{k})^{2-2\gamma\lambda_m} \end{array}$$

Sum all the two terms up,

$$\begin{array}{l} \sum_{k=1}^{n-1}||\frac{\gamma}{n} + \sum_{j=k}^{n-1}[\Pi_{i=j+1}^{n-1}(1 - \frac{\gamma A}{i+1})]\frac{(1-\gamma A)\gamma}{j(j+1)}||_2^2 \\ \leq 2\frac{\gamma^2}{n} + 2\gamma C_2(\frac{1}{n})(\frac{1}{n})^{2\gamma\lambda_m - 1}\sum_{k=1}^{n-1}(\frac{1}{k})^{2-2\gamma\lambda_m} \\ \leq 2\frac{\gamma^2}{n} + 2\gamma C_2(\frac{1}{n})(\frac{1}{n})^{2\gamma\lambda_m - 1}(\frac{1}{n})^{1-2\gamma\lambda_m} \\ \leq C_3\frac{1}{n} \end{array}$$

Now, define $M = \max\{||\pi_1 - \pi^*||_2^2, ||\Sigma_\omega||_2^2, ||\Sigma_d||_2^2\}$, we have

$$\begin{array}{rl} \sum_{i=1}^{n} L_n & = \mathbb{E}\sum_{i=1}^{n}||A(\pi_i - \pi^*)||_2^2 \\ & \leq \sum_{n=1}^{T}||A||_2^2[(C_2 + C_1)\frac{1}{n}]M \leq C\log(T). \end{array}$$

If $|\mathcal{D}|$ is finite, and we use a separate stochastic approximation to calculate the retail price, then the accumulated regret $\sum_{n=1}^{T} R_n \leq C|\mathcal{D}|\log(T)$.

■

Proof of Theorem 4

For a particular sequence of $d_1^{\text{DA}}, ...d_n^{\text{DA}}$, the regret term at day n has a linear form,

$$\begin{array}{rcl} R_n^\mu & = & \mathbb{E}||b - A\pi_n - d_n^{\text{DA}}||_2^2 \\ & = & \mathbb{E}||(1 - \sum_{i=1}^{n-1} P_i)b \\ & & -\sum_{t=1}^{n-1}(P_t w_t) + \sum_{t=1}^{n} Q_t d_t^{\text{DA}})||_2^2 \end{array}$$

Since b is arbitrary and in order to not introduce constant term, $1 - \sum_{i=1}^{n}\phi_i = 0$ should always hold. Therefore,

$$\begin{array}{rl} R_n^\mu & \geq \mathbb{E}\min_{\sum_{i=1}^{n-1} P_i = I}\sum_{i=1}^{n-1}||P_i||_2^2 \\ & \geq \frac{1}{n}\text{tr}(\Sigma_w^T\Sigma_w) \end{array}$$

Hence,

$$\sum_{n=1}^{T} R_n^\mu \geq C\log(T).$$

where $C = \text{tr}(\Sigma_w^T\Sigma_w)$. ■