An Online Learning Approach to Buying and Selling Demand Response [☆]

Kia Khezeli[a], Eilyan Bitar[a]

[a]*School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA.*

Abstract

We adopt the perspective of an aggregator, which seeks to coordinate its *purchase* of demand reductions from a fixed group of residential electricity customers, with its *sale* of the aggregate demand reduction in a two-settlement wholesale energy market. The aggregator procures reductions in demand by offering its customers a uniform price for reductions in consumption relative to their predetermined baselines. Prior to its realization of the aggregate demand reduction, the aggregator must also determine how much energy to sell into the two-settlement energy market. In the day-ahead market, the aggregator commits to a forward contract, which calls for the delivery of energy in the real-time market. The underlying aggregate demand curve, which relates the aggregate demand reduction to the aggregator's offered price, is assumed to be affine and subject to unobservable, random shocks. Assuming that both the parameters of the demand curve and the distribution of the random shocks are initially unknown to the aggregator, we investigate the extent to which the aggregator might dynamically adapt its offered prices and forward contracts to maximize its expected profit over a time window of T days. Specifically, we design a dynamic pricing and contract offering pol-

[☆]This work was supported in part by NSF grant ECCS-1351621, NSF grant IIP-1632124, US DoE under the CERTS initiative, and the Simons Institute for the Theory of Computing. This work builds on our preliminary results, presented at the IFAC 2017 World Congress (Khezeli et al., 2017). The current manuscript differs significantly from the conference version in terms of new results, formal proofs, and more detailed technical discussions. We thank Weixuan Lin for many helpful discussions, and his assistance in the proof of Lemma 2.

Email addresses: kk839@cornell.edu (Kia Khezeli), eyb5@cornell.edu (Eilyan Bitar)

icy that resolves the aggregator's need to learn the unknown demand model with its desire to maximize its cumulative expected profit over time. The proposed pricing policy is proven to be asymptotically optimal — exhibiting a *regret* over T days that is no greater than $O(\sqrt{T})$.

Keywords:
Demand response, dynamic pricing, online learning, electricity markets.

1. Introduction

The large scale utilization of demand response (DR) resources has the potential to substantially improve the reliability and efficiency of electric power systems. Accordingly, several state and federal mandates have been established to facilitate the integration of demand response resources into wholesale electricity markets. For example, FERC Order 719 mandates that Independent System Operators (ISOs) permit the direct sale of energy produced by DR resources into wholesale electricity markets (FERC, 2008). However, as individual residential customers often posses insufficient capacity to participate in such markets directly, there emerges the need for an intermediary, or *aggregator*, with the ability to coordinate the demand response of large numbers of residential customers for direct sale into the wholesale electricity market. Such is consistent with the growing multitude of ISO and utility-run DR programs, which require that aggregated DR resources have a minimum load curtailment capability of 100 kilowatt.[1]

In this paper, we adopt the perspective of an aggregator, which seeks to coordinate its *purchase* of an aggregate demand reduction from a fixed group of residential electricity customers, with its *sale* of the aggregate demand reduction into a two-settlement wholesale energy market.[2] Formally, this amounts to a two-sided optimization problem, which requires the aggregator to balance the cost it incurs in procuring a reduction in demand from participating customers against the revenue it derives from its sale of the (a

[1]Specific examples of such programs include the Proxy Demand Resource (PDR) program (Wolak et al., 2009) and the Day-Ahead Demand Response Program (DADRP) (NYISO, 2004) currently being operated by the California ISO and the New York ISO, respectively.

[2]From the perspective of the wholesale electricity market, the provisioning of a measurable reduction in demand from an aggregator is equivalent to an increase in supply.

priori uncertain) demand reduction into the wholesale energy market. We develop the problem more formally in what follows.

We consider the setting in which the aggregator purchases demand reductions from its customers using a non-discriminatory, posted price mechanism. That is to say, each participating customer is payed for her reduction in electricity demand according to a uniform per-unit energy price determined by the aggregator. Pricing mechanisms of this form fall within the more general category of DR programs that rely on peak time rebates (PTR) as incentives for demand reduction. Prior to its realization of the aggregate demand reduction, the aggregator must also determine how much energy to sell into the two-settlement energy market. In the day-ahead (DA) market, the aggregator commits to a forward contract, which calls for delivery of energy in the real-time (RT) market. If the realized reduction in demand exceeds (falls short of) the forward contract, then the difference is sold (bought) in the RT market. In order to maximize its profit, the aggregator must, therefore, co-optimize the DR price it offers its customers with the forward contract that it commits to in the wholesale energy market.

There are a variety of challenges that the aggregator faces in operating such DR programs. The most basic challenge is the prediction of how customers will adjust their aggregate demand in response to different DR prices, i.e., the aggregate demand curve. If the offered price is too low, consumers may be unwilling to curtail their demand; if the offered price is too high, the aggregator pays too much and gets more reduction than is needed. As the aggregator is initially ignorant to the customers' aggregate demand curve, the aggregator must attempt to learn a model of customer behavior over time through repeated observations of demand reductions in response to the DR prices that it offers. Simultaneously, the aggregator must jointly adjust its DR prices and forward contract offerings in such a manner as to facilitate profit maximization over time. As we will later show, such tasks are intimately related, and give rise to a trade-off between the need to *learn* (explore) and *earn* (exploit).

Contribution and Related Work: We study the setting in which the aggregator is faced with an aggregate demand curve that is affine in price, and subject to unobservable, additive random shocks. We assume that both the parameters of the demand curve and the probability distribution of the random shocks are fixed, and *initially unknown* to the aggregator. Faced with such ignorance, we explore the extent to which the aggregator might dynamically adapt its posted DR prices and offered contracts to maximize its expected profit over a time frame of T days. Specifically, we design a causal pricing and contract offering policy that resolves the aggregator's need to learn the unknown demand model with its desire to maximize its cumulative expected profit over time. The proposed pricing policy is proven to exhibit *regret* (relative to an oracle) over T days that is at most $O(\sqrt{T})$. In addition, the proposed policy generates a sequence of posted DR prices and forward contracts that converge to the oracle optimal DR price and forward contract in the mean square sense.

The literature — as it relates to the problem of co-optimizing an aggregator's decisions in both the retail and wholesale electricity markets — is sparse. Campaigne and Oren (2015) consider a market model that is perhaps closest in nature to the one considered in this paper. The authors adopt a mechanism design approach to eliciting demand response, where customers are rationed and remunerated according to their reported types. A related line of literature includes (Chao, 2012) and (Crampes and Léautier, 2015). In this paper, we take a posted price approach to the procurement of demand response. This is in sharp contrast to the mechanism design approach, as it gives rise to the need to learn customers' types (i.e., demand function) over time.

Organization: The remainder of the paper is organized as follows. In Section 2, we formulate the aggregator's profit maximization problem. In Section 3, we propose a recursive estimation scheme to facilitate the online learning of the underlying demand model. In Section 4, we propose a joint pricing and contract offering policy for the aggregator, and provide a theoretical analysis that establishes a sublinear growth rate of the regret incurred by the policy. In Section 5, we illustrate the performance of our proposed policy with a numerical example. Detailed proofs of all formal results can be found in the Appendix to the paper.

2. Model

We adopt the perspective of an aggregator who seeks to purchase demand reductions from a fixed group of N customers for sale into a two-settlement wholesale energy market. The market is assumed to repeat over multiple time periods (e.g., days) indexed by $t = 1, 2, \ldots$. The actions taken by the aggregator and their timing are specified in the following subsections.

2.1. Two-Settlement Market Model

At the beginning of each time period t, the aggregator commits to a forward contract for energy in the day-ahead (DA) market in the amount of Q_t (kWh). The forward contract is remunerated at the *DA energy price*. The forward contract calls for delivery in the real-time (RT) market. If the energy delivered by the aggregator (i.e., the aggregate demand reduction) falls short of the forward contract, the aggregator must purchase the shortfall in the RT market at the *shortage price*. If the energy delivered exceeds the forward contract, the aggregator must sell the excess supply in the RT market at the *overage price*. We denote the wholesale energy prices ($/kWh) by

- π, DA energy price,
- π_-, RT shortage price,
- π_+, RT overage price.

We assume throughout that the wholesale energy prices are fixed and known. We note, however, that it is straightforward to generalize the results stated in this paper to accommodate the more general setting in which the real-time energy prices are a priori uncertain and modeled as exogenous random variables (that are statistically independent of the underlying randomness in demand). We also assume that the wholesale energy prices satisfy $\pi > 0$ and $\pi_+ < \pi < \pi_-$. Such assumption serves to facilitate clarity of exposition and analysis in the sequel, as it preserves the concavity of the aggregator's expected profit function (2).

2.2. Demand Response Model

In order to fulfill its forward contract commitment Q_t on day t, the aggregator must elicit an aggregate reduction in demand from its customers. It does so by broadcasting a uniform DR price $p_t \geq 0$, to which each customer i responds with a reduction in demand in the amount of D_{it} (kWh), thereby entitling each customer i to receive a payment of $p_t D_{it}$. We note that implicit in this model is the assumption that each customer's reduction in demand is measured against a *predetermined baseline*. The question as to how to accurately estimate a customer's baseline consumption is a challenging and active area of research (Chao, 2011; Chelmis et al., 2015; Coughlin et al., 2009; Muthirayan et al., 2016). The expansion of our model to accommodate the endogenous estimation of a priori uncertain customer baselines is left as a direction for future research.

We model the response of each customer i to the posted price p_t at time t according to the *affine* function

$$D_{it} = a_i p_t + b_i + \varepsilon_{it}, \quad \text{for} \quad i = 1, \dots, N,$$

where $a_i \in \mathbb{R}$ and $b_i \in \mathbb{R}$ are customer i's idiosyncratic demand model parameters, and ε_{it} is an unobservable demand shock, which we model as a zero-mean random variable. We assume that both the model parameters a_i and b_i, and the probability distribution function of the demand shock are initially *unknown to the aggregator*. Clearly, the aggregate demand reduction $D_t := \sum_{i=1}^{N} D_{it}$ satisfies the affine relationship

$$D_t = ap_t + b + \varepsilon_t, \tag{1}$$

where the aggregate demand model parameters and shock are defined as $a := \sum_{i=1}^{N} a_i$, $b := \sum_{i=1}^{N} b_i$, and $\varepsilon_t := \sum_{i=1}^{N} \varepsilon_{it}$, respectively. In the sequel, we will occasionally denote the pair of aggregate demand parameters according to $\theta := (a, b)$.

We assume throughout the paper that $a \in [\underline{a}, \overline{a}]$ and $b \in \left[0, \overline{b}\right]$, where $\underline{a}$, $\overline{a}$, and $\overline{b}$ are assumed to be known and satisfy $0 < \underline{a} \leq \overline{a} < \infty$ and $0 \leq \overline{b} < \infty$. Such assumptions are natural, as they ensure a bounded and positive price elasticity of aggregate demand, and that reductions in aggregate demand are guaranteed to be nonnegative in the absence of demand shocks. We also assume that the sequence of aggregate demand shocks $\{\varepsilon_t\}$ are independent and identically distributed (IID) random variables, in addition to the following technical assumption.

Assumption 1. *The aggregate demand shock ε_t takes values in the interval $[\underline{\varepsilon}, \overline{\varepsilon}]$. Moreover, its cumulative distribution function F is bi-Lipschitz over this range. Namely, there exists a real constant $L \geq 1$, such that for all $x, y \in [\underline{\varepsilon}, \overline{\varepsilon}]$, it holds that*

$$\frac{1}{L}\,|x - y| \leq |F(x) - F(y)| \leq L\,|x - y|\,.$$

The assumption that the aggregate demand shock takes bounded values is natural, given the physical limitation on the range of values that demand can take. We also note that the aggregator does not require explicit knowledge of the parameters specified in Assumption 1 beyond the assumption of their boundedness.

2.3. Aggregator Profit

The expected profit derived by the aggregator during period t given a fixed price p_t and forward contract Q_t is determined by

$$r(p_t,Q_t) := \pi Q_t + \mathbb{E}\left[\pi_+[D_t - Q_t]^+ - \pi_-[Q_t - D_t]^+ - p_t D_t\right]. \qquad (2)$$

Here, expectation is taken with respect to the distribution on the random shock ε_t, and $[x]^+ := \max\{0, x\}$ for all $x \in \mathbb{R}$. It is not difficult to show that the expected profit criterion (2) is concave in its arguments (p_t, Q_t) given the assumptions stated in this paper thus far.

We define the *oracle optimal price and contract* as

$$(p^*, Q^*) := \text{argmax}\{r(p, Q) : (p, Q) \in \mathbb{R}^2\}.$$

That is to say, (p^*, Q^*) denote the DR price and forward contract, which jointly maximize the aggregator's expected profit given perfect knowledge of the demand model. Note that oracle optimal price and contract are time-invariant, as the wholesale energy prices and demand model are assumed to be time invariant. Their closed-form expressions are given in the following lemma.

Lemma 1 (Oracle Optimal Policy). *The oracle optimal price p^* and contract Q^* are given by*

$$p^* = \frac{1}{2}\left(\pi - \frac{b}{a}\right), \qquad (3)$$

$$Q^* = ap^* + b + F^{-1}(\alpha), \qquad (4)$$

where $\alpha := (\pi - \pi_+)/(\pi_- - \pi_+)$.

Here, $F^{-1}(\alpha) := \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}$ denotes the α-quantile of the random shock ε_t. We are guaranteed that $\alpha \in [0, 1]$, because of the assumption that $\pi_+ < \pi < \pi_-$. We define the *oracle optimal profit* accumulated over T time periods as

$$R^*(T) := \sum_{t=1}^{T} r(p^*, Q^*).$$

We employ the term *oracle*, as $R^*(T)$ equals the maximum expected profit that an aggregator might derive over T times periods if it had perfect knowledge of the demand model.

2.4. Policy Design and Regret

We consider the scenario in which the aggregator knows neither the demand model parameter $\theta = (a, b)$ nor the aggregate shock distribution F at the outset. Accordingly, the aggregator must endeavor to learn these features from the data it collects over time, e.g., through online assimilation of measurements of aggregate demand reductions in response to its posted DR prices. At the same time, the aggregator must dynamically adapt its sequence of posted DR prices (and forward contract offerings) to improve its profit over time. In what immediately follows, we describe the space of feasible policies that the aggregator might use to guide its adaptation of DR prices $\{p_t\}$ and contracts $\{Q_t\}$ over time.

Prior to its determination of the price p_t and the contract Q_t at time t, the aggregator has access to the entire history of prices, contract offerings, and aggregate demand reductions, up to and including time period $t-1$. We define a *feasible policy* as an infinite sequence of functions $\gamma := ((p_1, Q_1), (p_2, Q_2), \ldots)$, where each function in the sequence is allowed to depend only on the past data available until that point in time. More formally, we require that the functions (p_t, Q_t) be measurable according to the σ-algebra generated by the history of prices, offered contracts, and demand observations, i.e.,

$$(p_1, \ldots, p_{t-1}, Q_1, \ldots, Q_{t-1}, D_1, \ldots, D_{t-1})$$

for all time periods $t \geq 2$. For $t = 1$, we require that (p_1, Q_1) be a pair of deterministic constants.

The *expected profit* generated by a feasible policy γ over T time periods is defined as

$$R^\gamma(T) := \mathbb{E}^\gamma \left[\sum_{t=1}^{T} r(p_t, Q_t) \right], \tag{5}$$

where the expectation is taken with respect to the demand model (1) under the policy γ. We measure the performance of a feasible policy γ over T time periods according to the *T-period regret*:

$$\Delta^\gamma(T) := R^*(T) - R^\gamma(T).$$

The T-period regret incurred by a feasible policy equals the difference between the oracle optimal profit and the expected profit incurred by that policy over T time periods. Clearly, policies that produce low regret are preferred, as the oracle optimal profit is an upper bound on the maximum expected profit achievable by any feasible policy. Accordingly, we seek the design of policies whose T-period regret grows sublinearly with the horizon T. Such policies are said to have *no-regret*, as their average regret $(1/T) \cdot \Delta^\gamma(T)$ is guaranteed to vanish asymptotically. More formally, we have the following definition.

Definition 1 (No-Regret Policy). *A feasible policy γ is said to have* no-regret *if* $\lim_{T\to\infty} \Delta^\gamma(T)/T = 0$.

The following result establishes an upper bound on the T-period regret in terms of the pricing and contract errors relative to their oracle optimal counterparts. Lemma 2 will prove useful to the derivation of our main results.

Lemma 2. *The T-period regret incurred by any feasible policy γ is upper bounded by*

$$\Delta^\gamma(T) \le a \sum_{t=1}^{T} \mathbb{E}^\gamma\left[(p_t - p^*)^2\right] + L(\pi_- - \pi_+) \sum_{t=1}^{T} \mathbb{E}^\gamma\left[(Q_t - Q^* - a(p_t - p^*))^2\right], \tag{6}$$

where (p^*, Q^*) denote the oracle optimal price and contract.

Lemma 2 reveals that convergence of the posted prices $\{p_t\}$ and offered contracts $\{Q_t\}$ to the oracle optimal price p^* and contract Q^* in the mean square sense, respectively, will prove essential to the design of policies that exhibit no-regret. In the following section, we introduce a simple (least-squares) method for demand model learning that will facilitate the design of such policies.

3. Demand Model Learning

In this section, we propose a simple approach to enable the dynamic learning of the underlying demand model from data using the method of least squares estimation.

3.1. Parameter Estimation

We define the *least squares estimator* (LSE) of the parameter θ, given the history of past prices and demand observations through time period t as

$$\theta_t := \arg\min \left\{ \sum_{k=1}^{t} \left(D_k - (\vartheta_1 p_k + \vartheta_2)\right)^2 : (\vartheta_1, \vartheta_2) \in \mathbb{R}^2 \right\},$$

for time periods $t = 2, 3, \ldots$. The LSE is given by

$$\theta_t = \mathscr{J}_t^{-1} \left(\sum_{k=1}^{t} \begin{bmatrix} p_k \\ 1 \end{bmatrix} D_k \right), \tag{7}$$

assuming that the indicated inverse exists. The matrix $\mathscr{J}_t$ is defined as

$$\mathscr{J}_t := \sum_{k=1}^{t} \begin{bmatrix} p_k \\ 1 \end{bmatrix} \begin{bmatrix} p_k \\ 1 \end{bmatrix}^{\top}.$$

Its inverse is given by

$$\mathscr{J}_t^{-1} = \frac{1}{J_t} \left(\frac{1}{t} \sum_{k=1}^{t} \begin{bmatrix} -1 \\ p_k \end{bmatrix} \begin{bmatrix} -1 \\ p_k \end{bmatrix}^{\top} \right), \tag{8}$$

where $J_t := \sum_{k=1}^{t} (p_k - \bar{p}_t)^2$ and $\bar{p}_t := (1/t) \sum_{k=1}^{t} p_k$. Qualitatively J_t can be interpreted as measuring the cumulative dispersion of the sequence of posted prices around their mean. The parameter estimation error that results under the LSE (7) can be expressed as

$$\theta_t - \theta = \mathscr{J}_t^{-1} \left(\sum_{k=1}^{t} \begin{bmatrix} p_k \\ 1 \end{bmatrix} \varepsilon_k \right). \tag{9}$$

Recalling our previous assumption that the unknown parameter θ belongs to a closed and compact set given by $\Theta := [\underline{a}, \overline{a}] \times [0, \overline{b}]$, one can improve upon the LSE (7) by projecting θ_t onto the set Θ. More precisely, define the *truncated least squares estimator* (TLSE) as

$$\widehat{\theta}_t := \arg\min \left\{ \|\vartheta - \theta_t\|_2 : \vartheta \in \Theta \right\}. \tag{10}$$

It clearly holds that $\|\widehat{\theta}_t - \theta\| \leq \|\theta_t - \theta\|$. We have the following result, which establishes a general upper bound on the rate at which the TLSE converges to the true parameter in probability under any feasible policy.

Lemma 3 (Role of Price Dispersion). *Let γ be a feasible policy. There exist finite positive constants λ_1 and λ_2 such that*

$$\mathbb{P}^{\gamma}\{\|\widehat{\theta}_t - \theta\|_1 > \delta\} \;\leq\; 2\exp\left(-\lambda_1 \delta^2 t\right) \;+\; 2\mathbb{E}^{\gamma}\left[\exp\left(-\lambda_2 \delta^2 J_t\right)\right] \tag{11}$$

for all $\delta > 0$ and $t \geq 2$.

The parameter estimation error bound in (11) suggests a sufficient condition on the sequence of prices that guarantees consistency of the truncated least squares estimator. Namely, the parameter estimation error converges to zero in probability if the sequence of prices are such that their *cumulative dispersion J_t* diverges in probability. In Section 4, we propose a pricing policy that generates enough variation in the sequence of prices to ensure that J_t grows unbounded with probability one.

3.2. Quantile Estimation

We propose an approach to the recursive estimation of the unknown quantile function using the residuals generated by the truncated LSE (10). Define the sequence of *residuals* associated with the estimator $\widehat{\theta}_t$ as

$$\widehat{\varepsilon}_{k,t} := D_k - (\widehat{a}_t p_k + \widehat{b}_t), \quad \text{for } k = 1, \ldots, t. \tag{12}$$

Define their *empirical distribution* as

$$\widehat{F}_t(x) := \frac{1}{t}\sum_{k=1}^{t} \mathbb{1}\{\widehat{\varepsilon}_{k,t} \leq x\},$$

and their corresponding *empirical quantile function* as $\widehat{F}_t^{-1}(\alpha) := \inf\{x \in \mathbb{R} : \widehat{F}_t(x) \geq \alpha\}$. It will prove useful to the subsequent analyses to express the empirical quantile function in terms of the order statistics associated with the sequence of residuals. The *order statistics* associated with the sequence $\widehat{\varepsilon}_{1,t}, \ldots, \widehat{\varepsilon}_{t,t}$ are defined as a permutation of the sequence denoted by $\widehat{\varepsilon}_{(1),t}, \ldots, \widehat{\varepsilon}_{(t),t}$, where

$$\widehat{\varepsilon}_{(1),t} \leq \widehat{\varepsilon}_{(2),t} \leq \ldots \leq \widehat{\varepsilon}_{(t),t}.$$

With the order statistics of the residuals in hand, one can express the empirical quantile function as

$$\widehat{F}_t^{-1}(\alpha) = \widehat{\varepsilon}_{(i),t}, \tag{13}$$

where i is the unique index that satisfies $i - 1 < t\alpha \leq i$. It is not difficult to show that this index is given by $i = \lceil t\alpha \rceil$. Using Equation (13), the quantile estimation error can be linked to the parameter estimation error via the following inequality,

$$|\widehat{F}_t^{-1}(\alpha) - F^{-1}(\alpha)| \leq |F_t^{-1}(\alpha) - F^{-1}(\alpha)| + \left(1 + |p_{(i)}|\right) \|\widehat{\theta}_t - \theta\|_1, \qquad (14)$$

where $F_t^{-1}(\alpha)$ is defined as the empirical quantile function associated with sequence of demand shocks $\varepsilon_1, \ldots, \varepsilon_t$.

It follows from the inequality in (14) that consistency of the quantile estimator (13) depends on consistency of both the parameter estimator $\widehat{\theta}_t$ and the empirical quantile function $F_t^{-1}(\alpha)$. We establish consistency of the parameter estimator under our proposed policy in Lemma 4. Clearly, consistency of the empirical quantile function $F_t^{-1}(\alpha)$ does not depend on the particular policy being used. In Proposition 1, we establish a bound on the rate at which the sequence $\{F_t^{-1}(\alpha)\}$ converges to $F^{-1}(\alpha)$ in probability.

Proposition 1. *There exists a finite positive constant μ_1 such that*

$$\mathbb{P}\{|F_t^{-1}(\alpha) - F^{-1}(\alpha)| > \delta\} \leq 2\exp(-\mu_1 \delta^2 t) \qquad (15)$$

for all $\delta > 0$ and $t \geq 2$.

We omit a formal proof of Proposition 1, as it can be obtained as a direct consequence of Lemma 2 in Dvoretzky et al. (1956) using Assumption 1 stated in this paper.

4. Learning to Buy and Sell with No-Regret

In what follows, we build on the approach to demand model learning outlined in Section 3 to construct a pricing and contract offering policy, which is guaranteed to exhibit *no-regret*. In doing so, we establish in Theorem 1 a $O(\sqrt{T})$ upper bound on the T-period regret incurred under the proposed policy.

4.1. Myopic Policy (MP)

We first introduce a natural approach to pricing and contract offering, which combines the model learning scheme outlined in Section 3 with a *myopic* approach to pricing and contract offering. That is to say, at each time period t, the aggregator estimates the demand model parameters and quantile function according to $\widehat{\theta}_{t-1}$ and $\widehat{F}_{t-1}^{-1}(\alpha)$ defined in (10) and (13), respectively, and sets the price and forward contract according to

$$\widehat{p}_t = \frac{1}{2}\left(\pi - \frac{\widehat{b}_{t-1}}{\widehat{a}_{t-1}}\right), \tag{16}$$

$$\widehat{Q}_t = \widehat{a}_{t-1}\widehat{p}_t + \widehat{b}_{t-1} + \widehat{F}_{t-1}^{-1}(\alpha). \tag{17}$$

Under this myopic policy,[3] the aggregator treats its demand model estimates in each period as if they were correct, and ignores the impact of its choice of price on its ability to accurately estimate the demand model in future time periods. As discussed in Section 3.1, consistency of the parameter estimator is reliant upon sufficient dispersion in the underlying sequence of prices. However, under the myopic policy the sequence of prices may converge prematurely to a fixed price (that is different from the oracle optimal price). As a consequence, the sequence of parameter estimates may also converge to values different from the true model parameter. This phenomenon, also known as *incomplete learning*, is well-documented in the adaptive control literature (Borkar and Varaiya, 1982; Kumar and Varaiya, 2015; Lai and Robbins, 1982) and the revenue management literature (den Boer and Zwart, 2013; Keskin and Zeevi, 2014). In Section 5, we conduct a numerical case study that suggests the occurrence of incomplete learning under the myopic policy. We refer the reader to Figure 1(c) for a graphical illustration of incomplete learning under the myopic policy.

4.2. Randomly Perturbed Myopic Policy (RPMP)

To prevent the occurrence of incomplete learning, we propose a novel policy that is guaranteed to generate adequate price dispersion through application of random perturbations to the myopic policy. We refer to this policy as the *randomly perturbed myopic policy* (RPMP). It is defined as

$$p_t = \begin{cases} \widehat{p}_t, & \text{if } \xi_t = 0, \\ p_{t-1} + \rho, & \text{if } \xi_t = 1, \end{cases} \tag{18}$$

$$Q_t = \widehat{a}_{t-1}p_t + \widehat{b}_{t-1} + \widehat{F}_{t-1}^{-1}(\alpha), \tag{19}$$

[3] It is worth noting that, in the adaptive control theory literature, such myopic policies are more commonly known as *certainty equivalent* policies.

where $\xi_t \sim \mathrm{Ber}(\eta t^{-r})$ defines a sequence of independent Bernoulli random variables with $\mathbb{P}\{\xi_t = 1\} = \eta t^{-r}$. Here, the parameters $\eta \in (0, 1]$, $\rho \in (0, \infty)$, and $r \in [0, \infty)$ are user specified constants. The parameter η determines, in part, the probability that a perturbation is applied to the myopic price at any given time period, while the parameter ρ determines the magnitude of this perturbation. In this paper, we allow the parameters η and ρ to be arbitrary,[4] and investigate the role that the parameter r plays in controlling the rate at which the perturbation probability decays over time.

Ultimately, the parameter r must be designed to balance a delicate tradeoff between exploration and exploitation. On the one hand, the probability that a perturbation occurs should decay at a rate that is *slow enough* to generate sufficient price dispersion necessary to ensure consistent parameter estimation (cf. Lemma 3). On the other hand, this perturbation probability should decay at a rate that is *fast enough* to ensure that the (deliberate) pricing errors do not accumulate too rapidly. In Theorem 1, we establish an upper bound on the T-period regret that captures this tradeoff, and show that a perturbation probability $\mathbb{P}\{\xi_t = 1\} = O(t^{-1/2})$ is optimal in the sense that it minimizes the asymptotic order of our upper bound on regret.

4.3. A Bound on Regret

In what follows, we establish an upper bound on the T-period regret incurred by the randomly perturbed myopic policy. As part of our main result in Theorem 1, we also characterize the optimal 'decay rate' for the perturbation probability. We first present an upper bound on the rate at which the cumulative price dispersion J_t grows under the randomly perturbed myopic policy.

Lemma 4 (Price Dispersion under RPMP). *Let $\delta > 0$ and $t \geq 2$. There exists a finite positive constant λ_3 such that, under the randomly perturbed perturbed myopic policy (18) and (19), the cumulative price dispersion J_t satisfies*

$$\mathbb{E}^{\gamma}\left[\exp\left(-\delta J_t\right)\right] \leq \exp\left(-h(\delta) t^{1-r}\right),$$

where $h(\delta) := \lambda_3(1 - \exp(-\delta\rho^2/4))$.

[4]It is worth noting that the parameters ρ and η play a role in determining the finite-time performance of the randomly perturbed myopic policy (RPMP). However, the asymptotic order of regret incurred under the RPMP remains unchanged for any choice of $\rho > 0$ and $0 < \eta \leq 1$.

The combination of Lemmas 3 and 4 yields an upper bound on the rate at which the parameter estimation error converges to zero (in probability) under the randomly perturbed myopic policy for any dispersion parameter $r \in [0, 1)$.[5] This characterization of the parameter estimation error will play a central role in the proof of Theorem 1, which establishes an $O(T^r \vee T^{1-r})$ upper bound on the T-period regret incurred by the randomly perturbed myopic policy.

Theorem 1 (Sub-linear Regret). *Let $r \in (0, 1)$. There exist finite positive constants C_0, C_1, and C_2 such that the T-period regret incurred under the randomly perturbed myopic policy (18) and (19) is upper bounded by*

$$\Delta^\gamma(T) \leq C_0 + C_1 \log(T) + \left(\frac{C_2}{1-r}\right) T^{1-r} + \left(\frac{C_2}{r}\right) T^r$$

for all $T \geq 2$.

The structure of the upper bound on regret in Lemma 1 reveals an exploration-exploitation trade-off in choosing the dispersion parameter r. Specifically, the $O(T^r)$ term captures the component of revenue loss driven by the parameter estimation error; and the $O(T^{1-r})$ term captures the component of revenue loss driven by the deliberate pricing errors that are incurred when price perturbations are applied. A smaller (larger) value of the dispersion parameter r implies a greater tendency towards exploration (exploitation) in pricing under the RPMP. Clearly, this exploration-exploitation trade-off is balanced by setting the dispersion parameter to $r = 1/2$, as this value minimizes the asymptotic order of our upper bound on regret, yielding $\Delta^\gamma(T) \leq O(\sqrt{T})$.

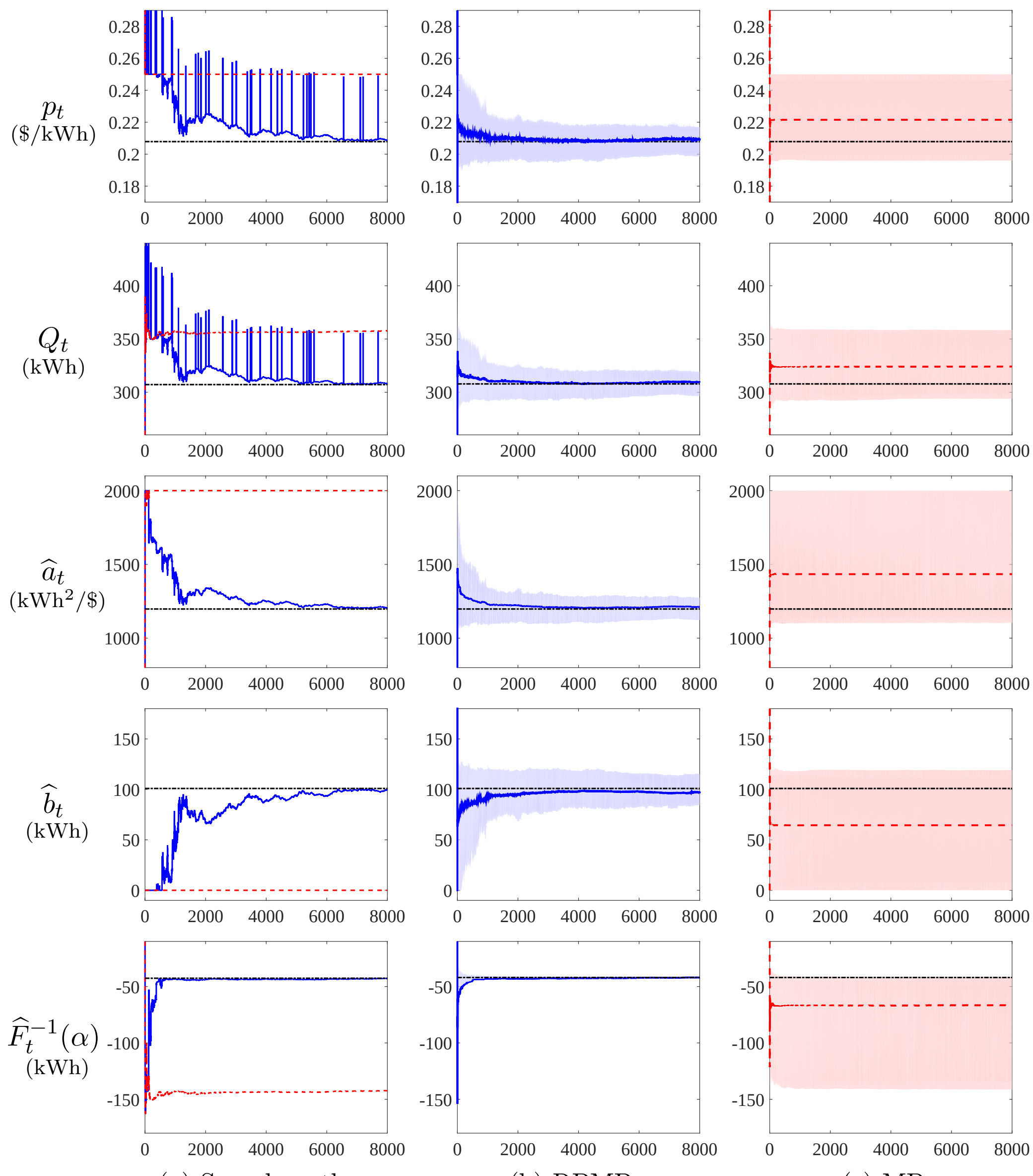
It is also worth noting that, as part of the proof of Theorem 1, we establish that the sequences of posted prices $\{p_t\}$ and contracts $\{Q_t\}$ generated by the randomly perturbed myopic policy converge to the oracle optimal price p^* and contract Q^* in the mean square sense, respectively. We also remark that Chen et al. (2014) consider a similar setting, which entails the online control of a dynamic inventory system through pricing and ordering decisions. They consider a different class of policy designs, and similarly establish an $O(\sqrt{T})$ upper bound on the order of regret for the class of policies they consider.

[5] We note that for dispersion parameters $r \geq 1$, the upper bound in Lemma 4 does not provide any useful information.

5. Numerical Case Study

We compare the performance of the myopic policy (MP) against the randomly perturbed myopic policy (RPMP) over a time horizon of $T = 10^4$ periods. We set the tuning parameters of the RPMP as $\eta = 0.2$, $\rho = 0.04$, and $r = 0.5$. This choice of ρ amounts to increasing the DR price offered to customers by four cents anytime a perturbation is applied. We assume that there are $N = 10^4$ customers participating in the DR program. For each customer i, we select a_i uniformly at random from the interval $[0.04, 0.20]$, and independently select b_i according to an exponential distribution (with mean equal to 0.01) truncated over the interval $[0, 0.1]$.[6] We further assume that the idiosyncratic demand parameters are drawn independently across customers. For each customer i, we let the demand shock have a normal distribution with zero-mean and standard deviation equal to 0.5, truncated over the interval $[-2, 2]$. We set the DA energy price, the RT shortage price, and the RT overage price to $\pi = 0.5$, $\pi_- = 1.7$, and $\pi_+ = 0.2$ (\$/kWh), respectively. Finally, we estimate the empirical means and confidence intervals associated with price, contract, and parameter estimate trajectories using 500 independent realizations of the experiment.

[6] This range of parameter values is consistent with the range of demand price elasticities observed in several real-time pricing programs operated in the United States (DoE, 2006; Faruqui and Sergici, 2010).



(a) Sample paths.

(b) RPMP.

(c) MP.

Figure 1: The set of figures on the left demonstrate sample paths generated by the *randomly perturbed myopic policy* (——), the *myopic policy* (- - -), and the *oracle optimal policy* (·········). The set of middle and right figures demonstrate the mean and confidence interval associated with sequences generated by the RPMP (middle) and the MP (right), compared against their oracle policy counterparts. The shaded area represents their middle 70% empirical confidence interval estimated using 100 independent experiments.

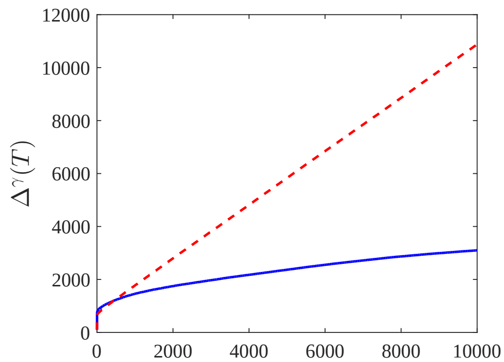


Figure 2: A plot of the T-period regret incurred by the randomly perturbed myopic policy (——) compared to the T-period regret incurred by the myopic policy (- - -).

5.1. Discussion

The plots in Figure 1(c) illustrate an apparent lack of exploration in the sequence of posted prices generated by the myopic policy. That is to say, the myopic price sequence rapidly converges to a fixed value, which on average differs substantially from the oracle optimal price. The same is true for the sequence of forward contracts generated by the myopic policy. The premature convergence of the myopic price sequence, in turn, leads to incomplete learning with the parameter estimates converging incorrect values. As a consequence, the T-period regret incurred by the myopic policy grows linearly in T, as shown in Figure 2.

On the other hand, the persistent variation in the sequence of prices generated by the randomly perturbed myopic policy induces parameter estimates, which asymptotically converge to the true parameter values, as can be seen from the plots in Figure 1(b). In particular, notice that the (middle 70%) empirical confidence intervals associated with the posted price and contract sequences generated by the randomly perturbed myopic policy shrink to their respective optimal oracle values over time. This provides empirical evidence supporting our theoretical claim that the sequences of prices and contracts generated by the randomly perturbed myopic policy converge to their oracle optimal values in probability.

6. Conclusion

In this paper, we study the problem of co-optimizing an aggregator's purchase and sale of demand response. The aggregator purchases energy in the form of demand reductions from a fixed group of residential customers, and sells the (a priori uncertain) aggregate demand reduction in a two-settlement wholesale electricity market. The customers' aggregate demand function is assumed to be affine in price (with unknown parameters) and subject to unobservable, additive random shocks (with unknown distribution). We propose a data-driven policy — referred to as the *randomly perturbed myopic policy* — to guide the aggregator's adaptation of its posted DR prices and forward contract offerings over time. We show that the proposed policy is consistent, meaning that the sequences of prices and contracts that it generates converge to the oracle optimal price and contract in the mean square sense, respectively. Moreover, we show that the regret incurred by the proposed policy over T time periods is no more than $O(\sqrt{T})$.

References

Borkar, V., Varaiya, P., 1982. Identification and adaptive control of markov chains. SIAM Journal on Control and Optimization 20 (4), 470–489.

Campaigne, C., Oren, S. S., 2015. Firming renewable power with demand response: an end-to-end aggregator business model. Journal of Regulatory Economics, 1–37.

Chao, H.-p., 2011. Demand response in wholesale electricity markets: the choice of customer baseline. Journal of Regulatory Economics 39 (1), 68–88.

Chao, H.-p., 2012. Competitive electricity markets with consumer subscription service in a smart grid. Journal of Regulatory Economics 41 (1), 155–180.

Chelmis, C., Saeed, M. R., Frincu, M., Prasanna, V., 2015. Curtailment estimation methods for demand response: Lessons learned by comparing apples to oranges. In: Proceedings of the 2015 ACM Sixth International Conference on Future Energy Systems. ACM, pp. 217–218.

Chen, B., Chao, X., Ahn, H.-S., 2014. Coordinating pricing and inventory replenishment with nonparametric demand learning. Available at SSRN 2694633.

Coughlin, K., Piette, M. A., Goldman, C., Kiliccote, S., 2009. Statistical analysis of baseline load models for non-residential buildings. Energy and Buildings 41 (4), 374–381.

Crampes, C., Léautier, T.-O., 2015. Demand response in adjustment markets for electricity. Journal of Regulatory Economics 48 (2), 169–193.

den Boer, A. V., Zwart, B., 2013. Simultaneously learning and optimizing using controlled variance pricing. Management science 60 (3), 770–783.

DoE, 2006. Benefits of demand response in electricity markets and recommendations for achieving them. US Dept. Energy, Washington, DC, USA, Tech. Rep.

Dvoretzky, A., Kiefer, J., Wolfowitz, J., 1956. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. The Annals of Mathematical Statistics, 642–669.

Faruqui, A., Sergici, S., 2010. Household response to dynamic pricing of electricity: a survey of 15 experiments. Journal of Regulatory Economics 38 (2), 193–225.

FERC, 2008. 719, Wholesale competition in regions with organized electric markets. Federal Energy Regulatory Commission.

Keskin, N. B., Zeevi, A., 2014. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. Operations Research 62 (5), 1142–1167.

Khezeli, K., Bitar, E., 2017a. An online learning approach to buying and selling demand response.
URL https://bitar.engineering.cornell.edu/papers/AE17KB.pdf

Khezeli, K., Bitar, E., 2017b. Risk-sensitive learning and pricing for demand response. IEEE Transactions on Smart Grid.

Khezeli, K., Lin, W., Bitar, E., 2017. Learning to buy (and sell) demand response. Proceedings of the IFAC World Congress.

Kumar, P. R., Varaiya, P., 2015. Stochastic systems: Estimation, identification, and adaptive control. SIAM.

Lai, T., Robbins, H., 1982. Iterated least squares in multiperiod control. Advances in Applied Mathematics 3 (1), 50–73.

Muthirayan, D., Kalathil, D., Poolla, K., Varaiya, P., 2016. Mechanism design for self-reporting baselines in demand response. In: American Control Conference (ACC), 2016. American Automatic Control Council (AACC), pp. 1446–1451.

NYISO, 2004. Day-ahead demand response program.
URL www.nyiso.com/public/webdocs/markets_operations/market_data/demand_response/Demand_Response/Day-Ahead_Demand_Response_Program/dadrp_primer.pdf

Wolak, F. A., Bushnell, J., Hobbs, B. F., 2009. The California ISO's proxy demand resource (PDR) proposal. Market Surveillance Committee of the California ISO May 1.

Appendix

7. Proof of Lemma 1

Given a fixed pair (p, Q), we have that

$$\begin{aligned} r(p,Q) =& \pi Q - (ap^2 + pb) + \pi_+ \mathbb{E}\left[ap + b - Q + \varepsilon_1\right]^+ \\ & - \pi_- \mathbb{E}\left[Q - (ap + b) - \varepsilon_1\right]^+ . \end{aligned}$$

It is straightforward to show that $r(p, Q)$ is strictly concave in its arguments. It follows that one can characterize its unique maximizers as solutions to the first order optimality conditions:

$$\frac{\partial r(p,Q)}{\partial p} = -2ap - b + a\pi_+(1 - F(Q - ap - b)) + a\pi_- F(Q - ap - b) = 0, \tag{20}$$

$$\frac{\partial r(p,Q)}{\partial Q} = \pi - \pi_+(1 - F(Q - ap - b)) - \pi_- F(Q - ap - b) = 0. \tag{21}$$

We note that implicit in calculation of the above partial derivatives is the exchange of the differentiation and integration operators. Under the operating assumptions of this paper, it is straightforward to prove the validity of such an exchange using the Dominated Convergence Theorem. The first order conditions (20) and (21) can be solved explicitly to obtain the desired result.

8. Proof of Lemma 2

Let $t \geq 1$, and fix (p_t, Q_t). To streamline the proof, we define $Y_t := Q_t - ap_t - b$ for each time period t. It follows that the expected profit of the aggregator can be expressed as

$$r(p_t, Q_t) = \pi Y_t + \mathbb{E}\left[\pi_+[\varepsilon_t - Y_t]^+ - \pi_-[Y_t - \varepsilon_t]^+\right] + (\pi - p_t)(ap_t + b).$$

It will be helpful to decompose the expected profit as $r(p_t, Q_t) = r_1(p_t, Q_t) + r_2(p_t, Q_t)$, where

$$\begin{aligned} r_1(p_t, Q_t) &:= \pi Y_t + \mathbb{E}\left[\pi_+[\varepsilon_t - Y_t]^+ - \pi_-[Y_t - \varepsilon_t]^+\right], \\ r_2(p_t, Q_t) &:= (\pi - p_t)(ap_t + b). \end{aligned}$$

First, it is straightforward to show that

$$r_2(p^*, Q^*) - r_2(p_t, Q_t) = a(p_t - p^*)^2. \tag{22}$$

Now we show that for each time period t, we have

$$r_1(p^*, Q^*) - r_1(p_t, Q_t) \leq L(\pi_- - \pi_+)(Y_t - Y^*)^2, \tag{23}$$

where $Y^* := Q^* - ap^* - b$. It is straightforward to show that $Y^* = F^{-1}(\alpha)$. Consider first the case in which $Y_t \geq Y^*$. It follows that

$$
\begin{aligned}
& r_1(p^*, Q^*) - r_1(p_t, Q_t) \\
&= \pi(Y^* - Y_t) + \pi_+ \int_{Y^*}^{\infty} (\varepsilon_t - Y^*) dF - \pi_+ \int_{Y_t}^{\infty} (\varepsilon_t - Y_t) dF \\
&\quad - \pi_- \int_{-\infty}^{Y^*} (Y^* - \varepsilon_t) dF + \pi_- \int_{-\infty}^{Y_t} (Y_t - \varepsilon_t) dF \\
&= \pi(Y^* - Y_t) + \pi_+ \int_{Y^*}^{+\infty} (Y_t - Y^*) \mathrm{d}F + \pi_- \int_{-\infty}^{Y^*} (Y_t - Y^*) \mathrm{d}F \\
&\quad + (\pi_- - \pi_+) \int_{Y^*}^{Y_t} (Y_t - \varepsilon_t) \mathrm{d}F \\
&= (Y^* - Y_t)\big(\pi - \pi_+(1 - F(Y^*)) - \pi_- F(Y^*)\big) + (\pi_- - \pi_+) \int_{Y^*}^{Y_t} (Y_t - \varepsilon_t) \mathrm{d}F \\
&= (\pi_- - \pi_+) \int_{Y^*}^{Y_t} (Y_t - \varepsilon_t) \mathrm{d}F,
\end{aligned}
$$

where the last equality follows from the fact that $F(Y^*) = F(F^{-1}(\alpha)) = \alpha$. Now, using the fact that F is bi-Lipschitz, we obtain

$$
\begin{aligned}
r_1(p^*, Q^*) - r_1(p_t, Q_t) &= (\pi_- - \pi_+) \int_{Y^*}^{Y_t} (Y_t - \varepsilon_t) \mathrm{d}F \\
&\leq (\pi_- - \pi_+)(Y_t - Y^*) \int_{Y_t^*}^{Y_t} \mathrm{d}F \\
&\leq L(\pi_- - \pi_+)(Y_t - Y^*)^2.
\end{aligned}
$$

For the case in which $Y_t < Y^*$, one can obtain an identical upper bound using an analogous approach as above. Finally, combining Inequality (23) and Equation (22) with the fact that $Y_t - Y^* = Q_t - Q^* - a(p_t - p^*)$ yields the desired upper bound on regret.

9. Proof of Lemma 3

Using Equation (9), it is straightforward to show that the parameter estimation error is bounded from above by

$$
\|\widehat{\theta}_t - \theta\|_1 \leq (1 + \overline{p})|\breve{\varepsilon}_t| + |\overline{\varepsilon}_t|, \tag{24}
$$

where $\bar{\varepsilon}_t$ and $\breve{\varepsilon}_t$ are defined as follows,

$$\bar{\varepsilon}_t := \frac{1}{t}\sum_{k=1}^{t}\varepsilon_k \quad \text{and} \quad \breve{\varepsilon}_t := J_t^{-1}\sum_{k=1}^{t}(p_k - \bar{p}_t)\varepsilon_k.$$

It follows that

$$\mathbb{P}^\gamma\{\|\widehat{\theta}_t - \theta\|_1 \geq \delta\} \leq \mathbb{P}\{|\bar{\varepsilon}_t| \geq \delta/2\} + \mathbb{P}^\gamma\{(1+\overline{p})|\breve{\varepsilon}_t| \geq \delta/2\}. \tag{25}$$

Note that the first probability in the right hand side of Inequality (25) is independent of policy γ. To bound each term in Inequality (25), we use the following result, which was first introduced in (Khezeli and Bitar, 2017b, Lemma 2).

Lemma 5. *Let $\{X_k\}$ be an infinite sequence of zero mean independent random variables, satisfying $-\infty < \underline{X} \leq X_k \leq \overline{X} < \infty$, almost surely, for all k. Let $\{\beta_k\}$ be an infinite sequence of real numbers, and define the sequence of random variables*

$$Y_t := \left(\sum_{k=1}^{t}\beta_k X_k\right) \Bigg/ \left(\sum_{k=1}^{t}\beta_k^2\right).$$

For all $\gamma > 0$ and $t \geq 1$, it holds that

$$\mathbb{P}\left\{Y_t \geq \delta\right\} \leq \exp\left(-\frac{2\gamma^2}{(\overline{X}-\underline{X})^2}\sum_{k=1}^{t}\beta_k^2\right).$$

Now, by setting $X_k = \varepsilon_k$ and $\beta_k = 1$, and applying Lemma 5, we get

$$\mathbb{P}\{|\bar{\varepsilon}_t| \geq \delta/2\} \leq 2\exp\left(-\frac{\delta^2 t}{2(\overline{\varepsilon}-\underline{\varepsilon})^2}\right). \tag{26}$$

Similarly, by setting $X_k = \varepsilon_k$ and $\beta_k = p_k - \bar{p}_t$, and applying Lemma 5, we get

$$\mathbb{P}^\gamma\{|\breve{\varepsilon}_t| \geq \delta/2\} \leq \mathbb{E}^\gamma\left[2\exp\left(-\frac{\delta^2}{2(\overline{\varepsilon}-\underline{\varepsilon})^2}J_t\right)\right]. \tag{27}$$

By applying Inequalities (26) and (27) to Inequality (25), we get

$$\begin{aligned}\mathbb{P}^\gamma\{\|\widehat{\theta}_t - \theta\|_1 \geq \delta\} \leq{}& 2\exp\left(-\frac{\delta^2 t}{2(\overline{\varepsilon}-\underline{\varepsilon})^2}\right)\\ &+\mathbb{E}^\gamma\left[2\exp\left(-\frac{\delta^2}{2(1+\overline{p})^2(\overline{\varepsilon}-\underline{\varepsilon})^2}J_t\right)\right].\end{aligned}$$

Defining $\lambda_1 := 1/(2(\overline{\varepsilon}-\underline{\varepsilon})^2)$ and $\lambda_2 := 1/(\lambda_1(1+\overline{p})^2)$ yields the desired inequality.

10. Proof of Lemma 4

Recall the definition of cumulative price dispersion $J_t := \sum_{k=1}^{t}(p_k - \overline{p}_t)^2$. Under the RPMP, we have that

$$\begin{aligned}
J_t &= \sum_{k=1}^{t}(p_k - \bar{p}_t)^2 \\
&= (p_1 - \bar{p}_t)^2 + \sum_{k=2}^{t}\left((\widehat{p}_k - \bar{p}_t)(1-\xi_k) + (p_{k-1} - \bar{p}_t + \rho)\xi_k\right)^2 \\
&\geq \sum_{k=2}^{t}\left((\widehat{p}_k - \bar{p}_t)(1-\xi_k) + (p_{k-1} - \bar{p}_t + \rho)\xi_k\right)^2 \\
&= \sum_{k=1}^{t-1}\left((\widehat{p}_{k+1} - \bar{p}_t)(1-\xi_{k+1}) + (p_k - \bar{p}_t + \rho)\xi_{k+1}\right)^2 \\
&\geq \sum_{k=1}^{t-1}(p_k - \bar{p}_t + \rho)^2\xi_{k+1}^2.
\end{aligned}$$

The last inequality follows from the fact that $\xi_k(1-\xi_k) = 0$ for all k. By adding and subtracting the term $\sum_{k=1}^{t-1}(p_k - \bar{p}_t)^2$ to the right hand side of the above inequality, we get

$$\begin{aligned}
J_t &\geq \sum_{k=1}^{t-1}(p_k - \bar{p}_t + \rho)^2\xi_{k+1}^2 \\
&= \sum_{k=1}^{t-1}\left\{(p_k - \bar{p}_t + \rho)^2\xi_{k+1}^2 + (p_k - \bar{p}_t)^2\right\} - \sum_{k=1}^{t-1}(p_k - \bar{p}_t)^2 \\
&= \sum_{k=1}^{t-1}\left\{\left(p_k - \bar{p}_t + \frac{\xi_{k+1}^2}{1+\xi_{k+1}^2}\rho\right)^2(1+\xi_{k+1}^2) + \frac{\xi_{k+1}^2}{1+\xi_{k+1}^2}\rho^2\right\} - \sum_{k=1}^{t-1}(p_k - \bar{p}_t)^2 \\
&\geq \sum_{k=1}^{t-1}\left\{\left(p_k - \bar{p}_t + \frac{\xi_{k+1}^2}{1+\xi_{k+1}^2}\rho\right)^2(1+\xi_{k+1}^2) + \frac{\xi_{k+1}^2}{1+\xi_{k+1}^2}\rho^2\right\} - J_t \\
&\geq \rho^2\sum_{k=1}^{t-1}\frac{\xi_{k+1}^2}{1+\xi_{k+1}^2} - J_t, \qquad (28)
\end{aligned}$$

where the second equality follows from the following algebraic identity. For any real numbers x, y, and z, it holds that

$$(x+y)^2 z^2 + x^2 = \left(x + \frac{z^2}{1+z^2} y\right)^2 (1+z^2) + \frac{z^2}{1+z^2} y^2.$$

It follows from Inequality (28) that

$$J_t \geq \rho^2 \sum_{k=2}^{t} \frac{\xi_k^2}{2(1+\xi_k^2)}. \tag{29}$$

Thus, under the RPMP, we have that

$$\begin{aligned}
\mathbb{E}^{\gamma}\left[\exp\left(-\delta J_t\right)\right] &\leq \mathbb{E}\left[\exp\left(-\delta\rho^2 \sum_{k=2}^{t} \frac{\xi_k^2}{2(1+\xi_k^2)}\right)\right] \\
&= \prod_{k=2}^{t} \mathbb{E}\left[\exp\left(-\frac{\delta\rho^2}{2} \frac{\xi_k^2}{1+\xi_k^2}\right)\right] \\
&= \prod_{k=2}^{t} \left(1 - \eta k^{-r} + \eta k^{-r} \exp\left(-\frac{\delta\rho^2}{4}\right)\right),
\end{aligned}$$

where the first equality follows from the fact that sequence of random variables $\{\xi_k\}$ are independent, and the second equality follows from the fact that $\xi_k \sim \mathrm{Ber}(\eta k^{-r})$ for all k.

Using the fact that $0 \leq \eta k^{-r}(1 - \exp(-\delta\rho^2/4)) \leq 1$, and that $e^{-x} \geq 1 - x$ for all $x \geq 0$, we conclude that

$$\begin{aligned}
\mathbb{E}^{\gamma}\left[\exp\left(-\delta J_t\right)\right] &\leq \prod_{k=2}^{t} \exp\left(-\eta k^{-r}\left(1 - \exp\left(-\frac{\delta\rho^2}{4}\right)\right)\right) \\
&= \exp\left(-\eta\left(\sum_{k=2}^{t} k^{-r}\right)\left(1 - \exp\left(-\frac{\delta\rho^2}{4}\right)\right)\right) \\
&\leq \exp\left(-\eta \log(3/2)\, t^{1-r}\left(1 - \exp\left(-\frac{\delta\rho^2}{4}\right)\right)\right). \qquad (30)
\end{aligned}$$

The last inequality follows from the fact that for $t \geq 2$, we have that

$$\begin{aligned}\sum_{k=2}^{t} k^{-r} &\geq \int_{2}^{t+1} x^{-r} dx \\ &= \frac{1}{1-r}\left((t+1)^{1-r} - 2^{1-r}\right) \\ &\geq \frac{1}{1-r}\left((3/2)^{1-r} - 1\right) t^{1-r} \\ &\geq \log(3/2)\, t^{1-r}.\end{aligned}$$

Th last inequality follows from the fact that $((3/2)^{1-r} - 1)/(1-r)$ is a decreasing function of r over $(0, 1)$, and its limit at $r = 1$ is equal to $\log(3/2)$. Setting $\lambda_3 := \eta \log(3/2)$ yields the desired inequality.

11. Proof of Theorem 1

We first establish a result that relates pricing and contracting errors under the RPMP to the parameter estimation error. Its proof is postponed to Appendix 12.

Lemma 6. *Under the randomly perturbed myopic policy (18) and (19), it holds that*

$$\mathbb{E}^{\gamma}\left[(p_t - p^*)^2\right] \leq k_1 \mathbb{E}^{\gamma}\left[\|\widehat{\theta}_{t-1} - \theta\|_1^2\right] + k_2 t^{-r}, \tag{31}$$

and

$$\mathbb{E}^{\gamma}\left[(Q_t - Q^* - a(p_t - p^*))^2\right] \leq k_3 \mathbb{E}^{\gamma}\left[\|\widehat{\theta}_{t-1} - \theta\|_1^2\right] + k_4 \frac{1}{t-1} + k_5 \frac{1}{\sqrt{t-1}}, \tag{32}$$

for all $t \geq 2$.

We combine Lemmas 2 and 6 to obtain

$$\begin{aligned}\Delta^{\gamma}(T) \leq & k_0 + (k_1 + k_3 L(\pi_- - \pi_+)) \sum_{t=2}^{T} \mathbb{E}^{\gamma}\left[\|\widehat{\theta}_{t-1} - \theta\|_1^2\right] \\ & + \sum_{t=2}^{T}\left\{k_2 t^{-r} + L(\pi_- - \pi_+)\left(k_4 \frac{1}{t-1} + k_5 \frac{1}{\sqrt{t-1}}\right)\right\},\end{aligned}$$

where $k_0 := a(p_1 - p^*)^2 + L(\pi_- - \pi_+)(Q_1 - Q^* - a(p_1 - p^*)^2)$.

Note that for the TLSE $\widehat{\theta}_t$, it holds that $\mathbb{P}\{\|\widehat{\theta}_t - \theta\|_1 \geq \delta\} = 0$ for all $\delta > \overline{\delta}$, where $\overline{\delta}$ is defined as

$$\overline{\delta} := \overline{a} - \underline{a} + \overline{b}. \tag{33}$$

Using the fact that $\|\widehat{\theta}_t - \theta\|_1^2$ is a non-negative random variable, for $t \geq 2$, we get

$$\begin{aligned}
\mathbb{E}^\gamma\left[\|\widehat{\theta}_t - \theta\|_1^2\right] &= \int_0^\infty \mathbb{P}^\gamma\{\|\widehat{\theta}_t - \theta\|_1^2 \geq \delta\}d\delta \\
&= \int_0^{\overline{\delta}} \mathbb{P}^\gamma\{\|\widehat{\theta}_t - \theta\|_1 \geq \sqrt{\delta}\}d\delta \\
&\leq \int_0^{\overline{\delta}} 2\exp\left(-\lambda_1 \delta t\right) d\delta \\
&\quad + \int_0^{\overline{\delta}} 2\exp\left(-\lambda_3 t^{1-r}\left(1 - \exp\left(-\lambda_2 \rho^2 \delta/4\right)\right)\right) d\delta.
\end{aligned} \tag{34}$$

The inequality follows from a combination of Lemmas 3 and 4.We now bound each integral in (34) separately. For the first term, we get

$$\int_0^{\overline{\delta}} \exp\left(-\lambda_1 \delta t\right) d\delta = -\frac{1}{\lambda_1 t}\exp\left(-\lambda_1 \delta t\right)\Big|_0^{\overline{\delta}} \leq \frac{1}{\lambda_1 t}. \tag{35}$$

For the second term, using integration by substitution for $u = \exp(-\lambda_2 \rho^2 \delta/4)$, we get

$$\begin{aligned}
&\int_0^{\overline{\delta}} \exp\left(-\lambda_3 t^{1-r}\left(1 - \exp\left(-\lambda_2 \rho^2 \delta/4\right)\right)\right) d\delta \\
&\quad = \int_{\exp(-\lambda_2 \overline{\delta}^2 \rho^2/4)}^1 \frac{4\exp\left(-\lambda_3 t^{1-r}\left(1-u\right)\right)}{\lambda_2 \rho^2 u} du \\
&\quad \leq \int_{\exp(-\lambda_2 \overline{\delta} \rho^2/4)}^1 \frac{4\exp\left(-\lambda_3 t^{1-r}\left(1-u\right)\right)}{\lambda_2 \rho^2 \exp(-\lambda_2 \overline{\delta} \rho^2/4)} du \\
&\quad = \frac{4\exp\left(-\lambda_3 t^{1-r}\left(1-u\right)\right)}{t^{1-r}\lambda_3 \lambda_2 \rho^2 \exp(-\lambda_2 \overline{\delta} \rho^2/4)}\Bigg|_{\exp(-\lambda_2 \overline{\delta} \rho^2/4)}^1 \\
&\quad \leq \frac{4}{t^{1-r}\lambda_3 \lambda_2 \rho^2 \exp(-\lambda_2 \overline{\delta} \rho^2/4)}.
\end{aligned} \tag{36}$$

By applying Inequalities (35) and (36) to (34), we get

$$\mathbb{E}^{\gamma}\left[\|\widehat{\theta}_t - \theta\|_1^2\right] \leq \frac{2}{\lambda_1 t} + \frac{8}{t^{1-r}\lambda_3\lambda_2\rho^2\exp(-\lambda_2\overline{\delta}\rho^2/4)}.$$

Now by applying the above bound on the parameter estimation error to Inequality (32), we get

$$\begin{aligned}
\Delta^{\gamma}(T) &\leq k_0 + (k_1 + k_3 L(\pi_- - \pi_+)) \sum_{t=2}^{T} \frac{2}{\lambda_1} \frac{1}{t-1} \\
&\quad + (k_1 + k_3 L(\pi_- - \pi_+)) \sum_{t=2}^{T} \frac{8}{\lambda_3\lambda_2\rho^2\exp(-\lambda_2\overline{\delta}\rho^2/4)} \frac{1}{(t-1)^{1-r}} \\
&\quad + \sum_{t=2}^{T} \left\{ k_2 t^{-r} + L(\pi_- - \pi_+) \left(k_4 \frac{1}{t-1} + k_5 \frac{1}{\sqrt{t-1}} \right) \right\} \\
&\leq C_0 + (k_1 + k_3 L(\pi_- - \pi_+)) \sum_{t=2}^{T} \left\{ \frac{2}{\lambda_1} \frac{1}{t} + \frac{8}{\lambda_3\lambda_2\rho^2\exp(-\lambda_2\overline{\delta}\rho^2/4)} \frac{1}{t^{1-r}} \right\} \\
&\quad + \sum_{t=2}^{T} \left\{ k_2 t^{-r} + L(\pi_- - \pi_+) \left(k_4 \frac{1}{t} + k_5 \frac{1}{\sqrt{t}} \right) \right\} \\
&\leq C_0 + C_1 \log(T) + \left(\frac{C_2}{1-r} \right) T^{1-r} + \left(\frac{C_2}{r} \right) T^{r}.
\end{aligned}$$

The second inequality follows from the fact that

$$\sum_{t=2}^{T} \frac{1}{t-1} = \sum_{t=1}^{T-1} \frac{1}{t} = 1 + \sum_{t=2}^{T-1} \frac{1}{t} \leq 1 + \sum_{t=2}^{T} \frac{1}{t}.$$

The last inequality follows from the fact that for $r \in (0, 1)$, we have that

$$\sum_{t=2}^{T} t^{-r} \leq \int_{1}^{T} x^{-r} dx \leq \frac{1}{1-r} T^{1-r},$$

and that

$$\sum_{t=2}^{T} \frac{1}{t} \leq \int_{1}^{T} \frac{1}{x} dx = \log(T).$$

We complete the proof by defining the constants C_0, C_1, and C_2 as follows.

$$\begin{aligned}
C_0 &:= k_0 + (k_1 + k_3 L(\pi_- - \pi_+)) \left(\frac{2}{\lambda_1} + \frac{8}{\lambda_3 \lambda_2 \rho^2 \exp(-\lambda_2 \overline{\delta} \rho^2/4)} \right) \\
&\quad + L(\pi_- - \pi_+)(k_4 + k_5) \\
C_1 &:= k_1 \frac{2}{\lambda_1} + L(\pi_- - \pi_+)) \left(\frac{2k_3}{\lambda_1} + k_4 \right) \\
C_2 &:= L(\pi_- - \pi_+) k_5 + \max \left\{ \frac{8(k_1 + k_3 L(\pi_- - \pi_+))}{\lambda_3 \lambda_2 \rho^2 \exp(-\lambda_2 \overline{\delta} \rho^2/4)} \ , \ k_2 \right\}
\end{aligned}$$

12. Proof of Lemma 6

Using the fact that $\xi_t(1-\xi_t) = 0$ for all $t \geq 2$, we obtain

$$\begin{aligned}
\mathbb{E}^\gamma \left[(p_t - p^*)^2 \right] &= \mathbb{E}^\gamma \left[((\widehat{p}_t - p^*)(1-\xi_t) + (p_{t-1} - p^* + \rho)\xi_t)^2 \right] \\
&= \mathbb{E}^\gamma \left[(\widehat{p}_t - p^*)^2 \right] \mathbb{E} \left[(1-\xi_t)^2 \right] + \mathbb{E}^\gamma \left[(p_{t-1} - p^* + \rho)^2 \right] \mathbb{E} \left[\xi_t^2 \right] \\
&= \mathbb{E}^\gamma \left[(\widehat{p}_t - p^*)^2 \right] (1 - \eta t^{-r}) + \mathbb{E}^\gamma \left[(p_{t-1} - p^* + \rho)^2 \right] \eta t^{-r} \\
&\leq \mathbb{E}^\gamma \left[(\widehat{p}_t - p^*)^2 \right] + \mathbb{E}^\gamma \left[(p_{t-1} - p^* + \rho)^2 \right] \eta t^{-r}.
\end{aligned}$$

The second equality follows from the fact that ξ_t is independent from $\widehat{p}_t$ and p_{t-1}. Using Equation (16), it is not difficult to show that

$$|\widehat{p}_t - p^*| \leq \frac{a+b}{2\underline{a}a} \|\widehat{\theta}_{t-1} - \theta\|_1.$$

Using the above inequality, and the fact that $|p_{t-1} - p^* + \rho| \leq \overline{p} + \rho$ for all $t \geq 2$, we obtain

$$\mathbb{E}^\gamma \left[(p_t - p^*)^2 \right] \leq k_1 \mathbb{E}^\gamma \left[\|\widehat{\theta}_{t-1} - \theta\|_1^2 \right] + k_2 t^{-r},$$

where

$$k_1 := \left(\frac{a+b}{2\underline{a}a} \right)^2 \quad \text{and} \quad k_2 := (\overline{p} + \rho)^2 \eta.$$

Under the randomly perturbed myopic policy for $t \geq 2$, we have that

$$\begin{aligned}
&\mathbb{E}^\gamma \left[(Q_t - Q^* - a(p_t - p^*))^2 \right] \\
&\qquad = \mathbb{E}^\gamma \left[\left((\widehat{\theta}_{t-1} - \theta) \begin{bmatrix} p_t \\ 1 \end{bmatrix} + \widehat{F}_{t-1}^{-1}(\alpha) - F^{-1}(\alpha) \right)^2 \right] \\
&\qquad \leq \mathbb{E}^\gamma \left[\left(2(1+\overline{p}) \|\widehat{\theta}_{t-1} - \theta\|_1 + |F_{t-1}^{-1}(\alpha) - F^{-1}(\alpha)| \right)^2 \right].
\end{aligned}$$

The inequality follows from Inequality (14). Recall that $\widehat{\theta}_t$ is the TLSE. Thus, it holds that $\|\widehat{\theta}_t - \theta\|_1 \leq \overline{\delta}$ surely, where $\overline{\delta}$ is defined in Equation (33). Thus,

$$\begin{aligned}\mathbb{E}^\gamma\left[(Q_t - Q^* - a(p_t - p^*))^2\right] &\leq 4(1+\overline{p})^2\mathbb{E}^\gamma\left[\|\widehat{\theta}_{t-1} - \theta\|_1^2\right] \\ &\quad + \mathbb{E}\left[(F_{t-1}^{-1}(\alpha) - F^{-1}(\alpha))^2\right] + 4\overline{\delta}(1+\overline{p})\mathbb{E}\left[|F_{t-1}^{-1}(\alpha) - F^{-1}(\alpha)|\right].\end{aligned}$$

To bound the second and the third terms in the right hand side of the above inequality, we use Inequality (15), and the fact that $|F_t^{-1}(\alpha) - F^{-1}(\alpha)|$ and $(F_t^{-1}(\alpha) - F^{-1}(\alpha))^2$ are non-negative random variables. For $t \geq 2$, we have that

$$\begin{aligned}\mathbb{E}\left[\left|F_t^{-1}(\alpha) - F^{-1}(\alpha)\right|\right] &= \int_0^\infty \mathbb{P}\left\{\left|F_t^{-1}(\alpha) - F^{-1}(\alpha)\right| \geq \delta\right\} d\delta \\ &\leq \int_0^\infty 2\exp\left(-\mu_1\delta^2 t\right) d\delta \\ &= \frac{\sqrt{\pi}}{\sqrt{\mu_1 t}}, \end{aligned} \tag{37}$$

and

$$\begin{aligned}\mathbb{E}\left[\left(F_t^{-1}(\alpha) - F^{-1}(\alpha)\right)^2\right] &= \int_0^\infty \mathbb{P}\left\{\left|F_t^{-1}(\alpha) - F^{-1}(\alpha)\right| \geq \sqrt{\delta}\right\} d\delta \\ &\leq \int_0^\infty 2\exp\left(-\mu_1\delta t\right) d\delta \\ &= \frac{2}{\mu_1 t}. \end{aligned} \tag{38}$$

We conclude the proof by defining the constants k_3, k_4, and k_5 as

$$k_3 := 4(1+\overline{p})^2, \quad k_4 := \frac{2}{\mu_1}, \quad \text{and} \quad k_5 := 4\overline{\delta}(1+\overline{p})\frac{\sqrt{\pi}}{\sqrt{\mu_1}}.$$