MANAGEMENT SCIENCE
Vol. 58, No. 3, March 2012, pp. 570–586
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1110.1426

Bayesian Dynamic Pricing Policies: Learning and Earning Under a Binary Prior Distribution

J. Michael Harrison, N. Bora Keskin
Graduate School of Business, Stanford University, Stanford, California 94305
{harrison_michael@gsb.stanford.edu, keskin_bora@gsb.stanford.edu}

Assaf Zeevi
Graduate School of Business, Columbia University, New York, New York 10027, assaf@gsb.columbia.edu

Motivated by applications in financial services, we consider a seller who offers prices sequentially to a stream of potential customers, observing either success or failure in each sales attempt. The parameters of the underlying demand model are initially unknown, so each price decision involves a trade-off between learning and earning. Attention is restricted to the simplest kind of model uncertainty, where one of two demand models is known to apply, and we focus initially on performance of the myopic Bayesian policy (MBP), variants of which are commonly used in practice. Because learning is passive under the MBP (that is, learning only takes place as a by-product of actions that have a different purpose), it can lead to incomplete learning and poor profit performance. However, under one additional assumption, a constrained variant of the myopic policy is shown to have the following strong theoretical virtue: the expected performance gap relative to a clairvoyant who knows the underlying demand model is bounded by a constant as the number of sales attempts becomes large.

Key words: revenue management; pricing; estimation; Bayesian learning; exploration–exploitation
History: Received January 15, 2010; accepted June 17, 2011, by Gérard P. Cachon, stochastic models and simulation. Published online in *Articles in Advance* October 14, 2011.

1. Introduction

We consider in this paper a problem that was first targeted for study at least 35 years ago, is relatively simple in structure, and is widely considered fundamental, but is also unsolved. Briefly stated, the problem is that of sequential pricing when the underlying demand model is unknown and the market response to any given price is confounded by statistical noise. In this situation the seller confronts a trade-off between exploration of the demand environment (learning) and expected immediate profit (earning).

1.1. Overview of the Problem and Approach

The particular variant of the learning-and-earning problem that we consider here has two salient features: (a) The seller offers prices sequentially to individual customers, observing either success or failure in each sales attempt. There is an underlying *demand model*, also called a *demand function* or *demand curve*, that gives the probability of success as a function of the price that is offered. (b) The parameters of the underlying demand model are fixed but initially uncertain, as opposed to problems where the demand model itself is changing over time.

Model feature (a) corresponds to what Phillips (2005) calls customized pricing; Phillips (2005, Chap. 11) lists a number of important application areas for customized pricing, including both business-to-business and business-to-consumer applications. One of those is the pricing of consumer credit, such as auto loans and credit card lending, where potential borrowers apply individually for loans, allowing banks to offer different prices to different applicants and compile uncensored data on successes and failures.

This situation offers obvious opportunities for price experimentation, but experimentation is surprisingly rare in practice (Phillips 2010); it is common to find banks that fix for months or even years the interest rate they charge for loans of a given type. Sales data allow the bank to estimate with reasonable accuracy the average demand rate under that particular interest rate, referred to hereafter as the *incumbent price*, but the bank remains uncertain about the average demand rates achievable with alternative prices.

Following common practice in the revenue management industry, we think in terms of a parametric model class (such as the logit family of demand functions, or the probit family, or the linear family) and adopt a Bayesian formulation, using a prior distribution over model parameters to express uncertainty about the demand environment. Also in conformity with standard practice, we take as our point of

departure the *myopic* policy, or greedy policy, that chooses each successive price to maximize immediate expected gain. Recognizing that it will probably be impossible to determine exactly optimal pricing policies, we are interested in suboptimal polices that have provably good properties and perform well in simulation studies.

We consider in our study a seller who knows the average demand rate under a given incumbent price. As noted earlier, that assumption is well motivated by common practice in commercial banking. If the underlying demand model comes from a two-parameter family like the linear, logit, or probit families, this means that the seller is uncertain about just one demand parameter; without loss of generality one can think of the unknown parameter as a measure of price sensitivity, such as demand elasticity at the incumbent price. Attention is restricted in this paper to the stylized problem where one of two demand models is known to apply, or to put that another way, the unknown model parameter is known to have one of two possible values. Because we adopt a Bayesian formulation, this means that attention is restricted to the case where the seller's prior distribution is concentrated on just two possible demand models, or two possible parameter values, referred to hereafter as the case of a *binary prior*.

Our motivation for considering the stylized binary-prior setting is the usual one: By analyzing carefully the simplest possible version of our target problem, we hope to shed light on the basic issues involved and gain insights applicable to more realistic settings. The binary-prior version of the learn-and-earn problem still defies exact solution, and readers will see that its analysis is surprisingly subtle.

1.2. Focus on Myopic Bayesian Policies

Bayesian inference has become the dominant approach to statistical estimation in companies that do analytical pricing (see Girard 2000, Preslan et al. 2005). Moreover, as reported by the Boston Consulting Group, pricing policies used in various industries are often myopic, putting no particular emphasis on active learning (Morel et al. 2003). A common practice is to first estimate unknown model parameters using Bayesian methods and then choose the optimal price given those parameter values (see Phillips 2005, Chap. 11; Talluri and van Ryzin 2004, Chap. 9). This conventional approach is not truly optimal because it ignores the fact that the estimate-and-optimize cycle will be repeated in the future; the conventional approach does not explicitly formulate a program of price experimentation nor does it explicitly address the trade-off between immediate earnings on the one hand and learning about demand parameters on the other. One representative of the conventional approach is what we call the myopic Bayesian policy, which chooses at each decision point the price that maximizes expected profit from the next sales opportunity, given the current (posterior) distribution over demand parameters, updating the posterior distribution as new price-response data accumulate. In this paper we aim to shed new light on what can go wrong under such a policy and how its deficiencies can be remedied.

1.3. Related Literature

In economics, probably the most influential and frequently cited work on learning-and-earning is that by Rothschild (1974), by Easley and Kiefer (1988), and by Aghion et al. (1991). In each case an infinite-horizon discounted formulation was employed, and the authors focused on the following question: Is it certain that a seller who follows an optimal policy will eventually obtain complete information about the underlying demand environment? Rothschild (1974) examined the case where the seller can choose prices from a finite set and showed that the answer is in general negative. That is, a seller who follows the optimal policy may never learn what demand model actually pertains, an outcome that McLennan (1984) later described as one of *incomplete learning*. Easley and Kiefer (1988) and Aghion et al. (1991) expanded on that finding by considering fairly general action spaces for the seller.

In the operations research and management science (OR/MS) realm, the term "revenue management" is commonly used to include tactical pricing problems of the kind considered here. To be more specific, we consider a problem of dynamic pricing, but without the inventory constraints that are usually included in OR/MS formulations (see Talluri and van Ryzin 2004, Chap. 5). Also, following a trend in recent research (see below), we do not treat demand parameters as known data but rather as a source of initial uncertainty that can be resolved via price experimentation.

Aviv and Pazgal (2005) were among the first OR/MS researchers to consider tactical pricing with such "model uncertainty." Their model involves a single unknown parameter that characterizes consumer demand, and in the interest of tractability, they assume a conjugate prior distribution for that parameter. Farias and van Roy (2010) is a recent paper of similar character, featuring a Bayesian formulation, a single unknown demand parameter, a conjugate prior distribution for that parameter, and reliance on dynamic programming methods. Farias and van Roy (2010) contains an up-to-date survey of OR/MS research on learning-and-earning, including an early paper by Lobo and Boyd (2003) that explores the idea of price experimentation for purposes of demand estimation.

A different approach, using a classical statistics framework without reliance on dynamic programming, is pursued by Besbes and Zeevi (2009). Their work treats both parametric model uncertainty and the case in which the demand model need not belong to any parametric family, so their work stands at the opposite extreme from studies assuming a single unknown demand parameter. Further connections to antecedent literature are also discussed in some detail by Besbes and Zeevi (2009). A common theme in the work of Farias and van Roy (2010) and Besbes and Zeevi (2009) is their emphasis on deriving suboptimal policies that have provably good performance. Our paper shares that theme but is otherwise quite distinct from both papers and from most other work in revenue management to date. The distinguishing feature of our study is its focus on the simple setting with a binary prior distribution, which allows a deeper analysis of the interplay between Bayesian learning and pricing, with particular emphasis on what we call myopic Bayesian policies. There has also been a recent interest in other variants of the learning-and-earning problem; e.g., Levina et al. (2009) present a machine learning approach to dynamic pricing with model uncertainty, analyzing the case of strategic consumers.

As indicated earlier, our work also intersects with the economics literature on price experimentation, initiated by Rothschild (1974). That pioneering study used the multiarmed bandit paradigm first formalized by Robbins (1951), which is one of the classical formulations of dynamic optimization under model uncertainty; it has been used extensively in a variety of fields (see Gittins 1989). The Bayesian formulation and dynamic programming methods employed by Rothschild (1974) are similar to what one sees in the revenue management literature discussed above, although few OR/MS papers seem to recognize that connection. There has been a modest, sporadic stream of research in economics that builds on the foundation laid by Rothschild (1974), including the influential work by Easley and Kiefer (1988) and by Aghion et al. (1991) cited earlier. In particular, there have been extensions in the direction of strategic experimentation, involving multiple firms rather than a single monopolist (see Bolton and Harris 1999) and experimentation when the demand model may change over time (see Keller and Rady 1999).

At least superficially, the antecedent paper most closely related to ours is that by McLennan (1984). He too considers the problem of sequential pricing with probabilistic response and unknown demand parameters and adopts a Bayesian formulation where one of two possible demand models is assumed to pertain, but he restricts attention to the case of linear demand models. Like Rothschild (1974), McLennan (1984) adopts an infinite-horizon formulation with discounting (the discounting is crucial). Although optimal policies cannot be calculated explicitly, he shows that if his model parameters satisfy certain restrictions, then incomplete learning can occur under an optimal policy. Thus, McLennan (1984) produces a family of examples that extend the negative finding of Rothschild (1974) to a model class with a continuum of available prices. Incomplete learning also plays a central role in our analysis, but the emphasis here is quite different from McLennan's: He shows that incomplete learning may be the seller's *optimal* choice in a model with discounting, whereas we show that it may occur, and often does occur, as a consequence of *myopic* decision making, or greedy decision making; see further discussion in §8.

1.4. Overview of the Analysis and Main Results

Section 2 describes in mathematical terms the problem to be addressed, including a definition of the Bayesian pricing policies to which we restrict attention. There we also prove some preliminary results and introduce the myopic Bayesian policy (MBP). Section 3 presents two illuminating numerical examples, and in §4 we undertake a theoretical analysis of the MBP, showing that it can and often does lead to incomplete learning.

Section 5 presents a key technical result regarding what we call "discriminative policies," and then in §§6 and 7 we consider variants of the MBP that are intuitively appealing and perform well both in theory and in simulation experiments. Our findings are different depending on whether the seller's incumbent price is or is not optimal under one of the two possible demand models. If it is not, then a simple kind of policy that we call *constrained* MBP (CMBP) has the following strong theoretical virtue: the expected performance gap relative to a clairvoyant who knows the underlying demand model is bounded by a constant as the number of sales attempts becomes large. This is the main result of §6, and to the best of our knowledge, it is stronger than any other performance bound in the literature of dynamic pricing. CMBP is not a single policy but rather a family of policies whose members are distinguished by what we call a *constraint parameter*, which can be tuned (using simulation) to optimize performance.

The simple form of a CMBP policy is tailored to the binary-prior setting; it is more or less obvious that such policies will not perform well with a general prior distribution. In §7 we introduce a generalized form of CMBP that involves a *sequence* of constraint parameters. We conjecture that policies of this form will perform well in a relatively general setting if the parameters are suitably chosen. In particular, it is necessary in a general setting to have the constraint parameters decrease toward zero as the number of

sales attempts becomes large. To build confidence in the generalized CMBP family, the following result is proved in §7: If the constraint parameters decrease to zero at a suitable rate, then one can still achieve an expected performance gap that is bounded by a constant, assuming again that the seller's incumbent price is not optimal under either of the two possible demand models.

For the more subtle and difficult case where the seller's incumbent price *is* optimal under one of the two possible demand models, we show at the end of §6 that no policy can achieve an expected performance gap that is bounded by a constant. Our feeling is that the difficulties arising in the analysis of this case are representative of what one encounters with a general (dispersed) prior for the unknown model parameter, and we do not attempt a complete treatment in this paper. Section 8 summarizes the main contributions of this paper, especially in comparison with previous work. The proofs of all formal results are postponed to a sequence of appendices that conclude the paper.

2. Problem Formulation and Preliminaries

2.1. Basic Model Elements

Consider a firm, hereafter called *the seller*, that offers a single product for sale to customers who arrive in sequential fashion. As a matter of convention, we associate with each successive customer a distinct sales "period" so that, for example, the phrase "period-t revenue" simply means revenue realized from the t^{th} arriving customer. In each period $t = 1, 2, \ldots$, the seller must choose a price p_t from a given interval $[l, u]$, where $0 \le l < u < \infty$, after which the seller experiences either success (a sale at the offered price p_t) or failure (no sale). The probability of success when the seller offers price p in any given period is $\rho(p)$; we call $\rho(\cdot)$ the ambient *demand model*. The marginal cost of the product being sold is set to zero without loss of generality (because prices can always be expressed as increments above cost); given this normalization, the terms "profit" and "revenue" can and will be used interchangeably.

Before the first customer arrives, nature chooses either $\rho_0(\cdot)$ or $\rho_1(\cdot)$ as the ambient demand model; this choice is not observed by the seller, and it remains fixed over the entire selling horizon. We encode this choice via the random variable

$$\chi = \begin{cases} 1 & \text{if } \rho(\cdot) = \rho_1(\cdot), \\ 0 & \text{if } \rho(\cdot) = \rho_0(\cdot), \end{cases} \tag{1}$$

and denote by q_0 the *prior probability* assigned by the seller to the event $\{\chi = 1\}$; this number is part of our problem data. (The subscript in the notation q_0 differentiates this initial probability assessment from the ones formed later, after sales outcomes are observed.) To exclude trivial cases, in which the seller knows the ambient demand model with certainty, we assume that $0 < q_0 < 1$. We shall occasionally refer to the event or condition $\{\chi = i\}$ as *hypothesis* i. If price p is chosen in a given period, then the seller's expected revenue in that period under hypothesis i is

$$r_i(p) := p\rho_i(p) \quad \text{for } i = 0, 1. \tag{2}$$

The only random variables other than χ that will figure in the development to follow are indicator variables $X_1, X_2, \ldots$ defined as follows: $X_t = 1$ if there is a sale (success) in period t, and $X_t = 0$ otherwise. Defining $X := (X_1, X_2, \ldots)$, we call X the *sales sequence*.

The seller is assumed to know the success probability $\hat{\rho}$ for an *incumbent price* $\hat{p} \in (l, u)$. Consistent with this assumption, we restrict attention to demand hypotheses that satisfy $\rho_i(\hat{p}) = \hat{\rho}$ for $i = 0, 1$. The demand models $\rho_0(\cdot)$ and $\rho_1(\cdot)$ are also assumed to be continuously differentiable and strictly decreasing over $[l, u]$, and we define the associated price elasticity functions $\varepsilon_i(\cdot)$ as usual (here and later, a prime denotes a derivative):

$$\varepsilon_i(p) := -\frac{p\rho_i'(p)}{\rho_i(p)} \quad \text{for } i = 0, 1. \tag{3}$$

Both $\varepsilon_0(\cdot)$ and $\varepsilon_1(\cdot)$ are assumed to be strictly increasing, from which it follows that each of the single-period expected revenue functions $r_i(\cdot)$ has a unique maximizer p_i^* in $[l, u]$. We assume that p_0^* and p_1^* are interior points of the feasible price range $[l, u]$ and without loss of generality that $p_0^* < p_1^*$. The first-order conditions for optimality then give the following:

$$\varepsilon_0(p_0^*) = \varepsilon_1(p_1^*) = 1, \quad \text{where } l < p_0^* < p_1^* < u. \tag{4}$$

We assume throughout that the incumbent price $\hat{p}$ lies in the interval $[p_0^*, p_1^*]$ because the opposite case is essentially trivial. (If the demand curves $\rho_0(\cdot)$ and $\rho_1(\cdot)$ do not intersect within that interval, then the arguments to follow can easily be modified to show that the myopic Bayesian policy itself gives excellent results, and so the issues addressed in this paper simply do not arise.) Monotonicity of the price elasticity functions then has the following implication.

Proposition 1. *The unique price $p \in [p_0^*, p_1^*]$ satisfying $\rho_0(p) = \rho_1(p)$ is $p = \hat{p}$.*

A price p that satisfies $\rho_0(p) = \rho_1(p)$ is *uninformative*, providing no information about the ambient demand model, so Proposition 1 can be verbally paraphrased as follows: Under our assumptions, the *unique* uninformative price between p_0^* and p_1^* is the incumbent price $\hat{p}$.

2.2. Pricing Policies, Induced Probabilities, and Posterior Beliefs

In our Bayesian formulation of the dynamic pricing problem, a *policy* is formally defined as a sequence $\pi = (\pi_1, \pi_2, \ldots)$, where each component π_t is a function that maps $[0,1] \to [l,u]$; the meaning or interpretation of the component functions π_t will become clear shortly. For each policy π and each realization of the sales sequence X, we define the associated prices $(p_1, p_2, \ldots)$ and posterior probabilities $(q_1, q_2, \ldots)$ through the following recursive procedure. In each successive period $t = 1, 2, \ldots$, set $p_t = \pi_t(q_{t-1})$ and then compute q_t using Bayes' rule:

$$q_t = \begin{cases} \dfrac{q_{t-1}\rho_1(p_t)}{q_{t-1}\rho_1(p_t) + (1-q_{t-1})\rho_0(p_t)} & \text{if } X_t = 1, \\ \dfrac{q_{t-1}[1-\rho_1(p_t)]}{q_{t-1}[1-\rho_1(p_t)] + (1-q_{t-1})[1-\rho_0(p_t)]} & \\ & \text{otherwise} \end{cases}$$

$$= \frac{q_{t-1}\rho_1(p_t)^{X_t}[1-\rho_1(p_t)]^{1-X_t}}{q_{t-1}\rho_1(p_t)^{X_t}[1-\rho_1(p_t)]^{1-X_t} + (1-q_{t-1})\rho_0(p_t)^{X_t}[1-\rho_0(p_t)]^{1-X_t}}. \tag{5}$$

One interprets q_t as the probability assigned by the seller to the event $\{\chi = 1\}$ after the first t sales outcomes have been observed; for brevity, we call q_t the seller's *belief* in hypothesis 1 after t periods.

A pricing policy π induces two probability measures $\mathbb{P}_0^\pi$ and $\mathbb{P}_1^\pi$ on the outcome space of X (that is, the space whose elements are sequences of zeros and ones) via the following formula:

$$\mathbb{P}_i^\pi(X_1 = x_1, \ldots, X_T = x_T) = \prod_{t=1}^{T} [\rho_i(p_t)]^{x_t}[1-\rho_i(p_t)]^{1-x_t}, \tag{6}$$

where $p_1, p_2, \ldots$ is the price sequence associated with π and the binary-valued sales realizations $x_1, x_2, \ldots$. One interprets $\mathbb{P}_i^\pi(A)$ as the probability of event A under policy π and demand hypothesis i. In the usual way, we denote by $\mathbb{E}_i^\pi(\cdot)$ the expectation operator associated with the probability measure $\mathbb{P}_i^\pi(\cdot)$.

Now, from the definition (5) it is easy to verify that the corresponding posterior probabilities $\{q_t\}$ form a bounded nonnegative supermartingale under $\mathbb{P}_0^\pi$ and form a bounded nonnegative submartingale under $\mathbb{P}_1^\pi$. (Here and later, when we make reference to martingales, the associated filtration is that generated by the sales indicators $X_1, X_2, \ldots$.) Thus, standard results in martingale theory can be used to establish the following conclusion.

Proposition 2 (Convergence of Beliefs). *For each pricing policy π, the posterior probabilities $\{q_t\}$ converge almost surely as $t \to \infty$ to a limit belief q_∞ under both $\mathbb{P}_0^\pi$ and $\mathbb{P}_1^\pi$.*

2.3. Performance Metrics

Again denoting by $p_1, p_2, \ldots$ the price sequence associated with a given policy π, we define the conditional expected revenue totals

$$R_i^\pi(T) = \mathbb{E}_i^\pi\left\{\sum_{t=1}^{T} r_i(p_t)\right\} \quad \text{for } i = 0, 1 \text{ and } T = 1, 2, \ldots, \tag{7}$$

where the expectation is taken over the sales indicators $X_1, \ldots, X_T$ that determine the prices $p_1, \ldots, p_T$. In the development to follow we focus primarily on the following performance metric:

$$\Delta_i^\pi(T) = \frac{1}{r_i(p_i^*)}[Tr_i(p_i^*) - R_i^\pi(T)] \quad \text{for } i = 0, 1 \text{ and } T = 1, 2, \ldots. \tag{8}$$

To understand the meaning of this quantity, note the following: A clairvoyant who knows which demand model actually applies will choose price p_i^* in every period when $\chi = i$, so the first term inside the square brackets on the right-hand side of (8) is the clairvoyant's expected T-period revenue under hypothesis i. Thus, the quantity in the square brackets is positive and expresses the expected T-period profit performance of policy π relative to what the clairvoyant would achieve, given that $\chi = i$. The definition (8) re-expresses that performance differential as a multiple of the clairvoyant's average profit per period; that is, $\Delta_i^\pi(T)$ is the number of periods of ideal expected profit that the seller loses over the first T periods because of demand model uncertainty, given that hypothesis i pertains and that the seller has chosen policy π. Because $\Delta_i^\pi(T)$ is nondecreasing as a function of T, the limit $\Delta_i^\pi(\infty)$ necessarily exists, although it may be infinite. Let us also define the performance measure $\Delta(\cdot) := \frac{1}{2}\Delta_0(\cdot) + \frac{1}{2}\Delta_1(\cdot)$, where $\Delta_i(\cdot)$ for $i \in \{0,1\}$ is given as in (8), recalling that the subscript i means that demand hypothesis i is true.

2.4. The Myopic Bayesian Policy (MBP)

If the seller enters a period with posterior belief q and chooses price p, then from the seller's perspective that period's expected profit is

$$r_q(p) := qr_1(p) + (1-q)r_0(p). \tag{9}$$

Define the *myopic price function*

$$\varphi(q) := \sup\{\arg\max\{r_q(p), l \le p \le u\}\} \quad \text{for } 0 \le q \le 1. \tag{10}$$

The use of the supremum in the above definition is necessary because for some values of q the function $r_q(\cdot)$ may achieve its maximum at multiple prices; readers will see shortly a numerical illustration of this observation. Now, recall that in §2 we defined $p_0^* := \varphi(0)$ and $p_1^* := \varphi(1)$, observing that the maximizing price in (10) *is* unique if $q = 0$ or $q = 1$.

Proposition 3. *The function $\varphi(\cdot)$ is nondecreasing on* $[0, 1]$.

The MBP is the pricing policy π that chooses $\pi_t(q) = \varphi(q)$ for all $t = 1, 2, \ldots$ and $q \in [0, 1]$. That is, the MBP puts all of its emphasis on earning, ignoring the trade-off that was highlighted in §1.

2.5. The Main Question Addressed in This Paper

Because beliefs are guaranteed to converge under any Bayesian policy (Proposition 2), one might plausibly hope that the limit belief under the MBP will correctly identify the ambient demand model; i.e., q_∞ will be equal to 1 if $\chi = 1$ and will be equal to 0 otherwise. In other words, even with a myopic focus on earning, learning will somehow "take care of itself." In the next section we will see that this hope is unfounded in general, with obvious negative consequences on the revenue performance of the MBP. Understanding and "fixing" that undesirable feature of the MBP is the main subject of this paper.

3. Two Motivating Examples

Figure 1 pictures $\rho_0(\cdot)$ and $\rho_1(\cdot)$ for two illuminating examples: In the first one, both demand models have the linear form $\rho(p) = a - bp$; in the second one, they both have the logit form $\rho(p) = [1 + \exp(a + bp)]^{-1}$. To be specific, the data for the linear example are

$$l = 0.5, \quad u = 1.5, \quad \rho_0(p) = 1.4 - 0.9p, \quad \text{and} \quad \rho_1(p) = 0.8 - 0.3p, \tag{11}$$

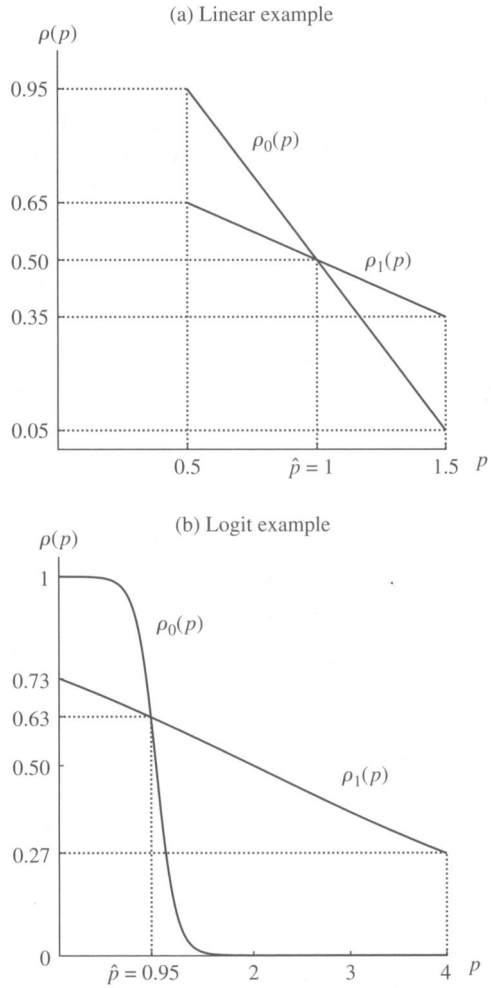
and for the logit example they are

$$l = 0, \quad u = 4, \quad \rho_0(p) = \frac{1}{1 + e^{-10+10p}}, \quad \text{and} \quad \rho_1(p) = \frac{1}{1 + e^{-1+0.5p}}. \tag{12}$$

Note that in each of these examples the incumbent price $\hat{p}$ is the unique uninformative price (that is, $\hat{p}$ is the unique price at which the two curves cross).

Figure 2(a) plots estimates of $\Delta(T)$, the performance function defined in the previous section, for the linear example (11) and various horizon lengths T. These estimates are based on 10,000 independently generated sales sequences, taking the seller's prior belief to be $q_0 = 0.5$. Figure 2(b) displays the strikingly different MBP performance in our logit example (12).

Figure 1 Two Examples of Demand Models from Standard Parametric Families

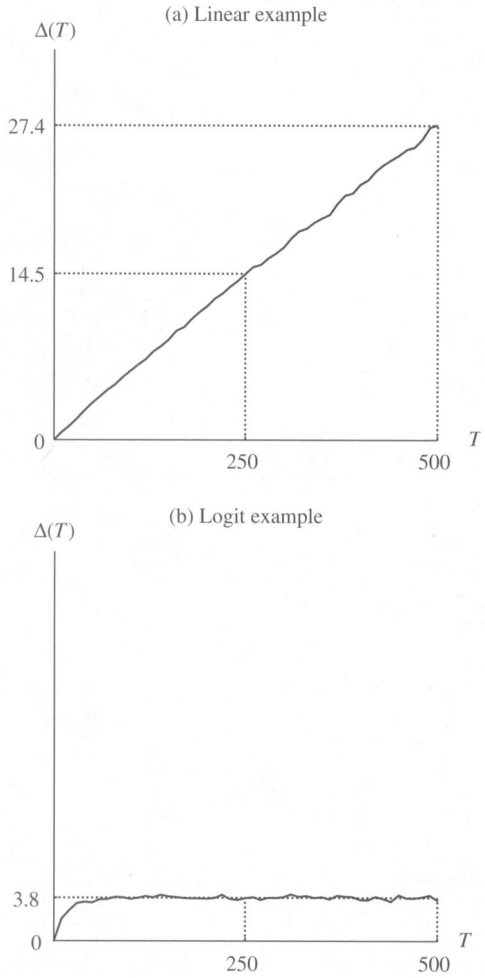


Notes. Panel (a) depicts two linear demand models, each of which specifies the probability of a successful sale at different price levels. Panel (b) shows two logit demand models.

To understand what is driving these numerical findings, we first note that in the linear example, the belief process under MBP fails to reveal the ambient demand model for more than half of the simulated scenarios. This, in conjunction with the definition of the performance function $\Delta(\cdot)$, implies the expected revenue loss must increase linearly, as indeed one observes in Figure 2(a). In our logit example, however, the MBP correctly identifies the ambient demand model, and this is achieved over a relatively short span of observed purchase decisions (see Figure 3). This is a favorable property, of course, but it does not fully explain the strikingly good performance seen in Figure 2(b), which suggests expected revenue loss is bounded by a *constant*, independent of the sales horizon.

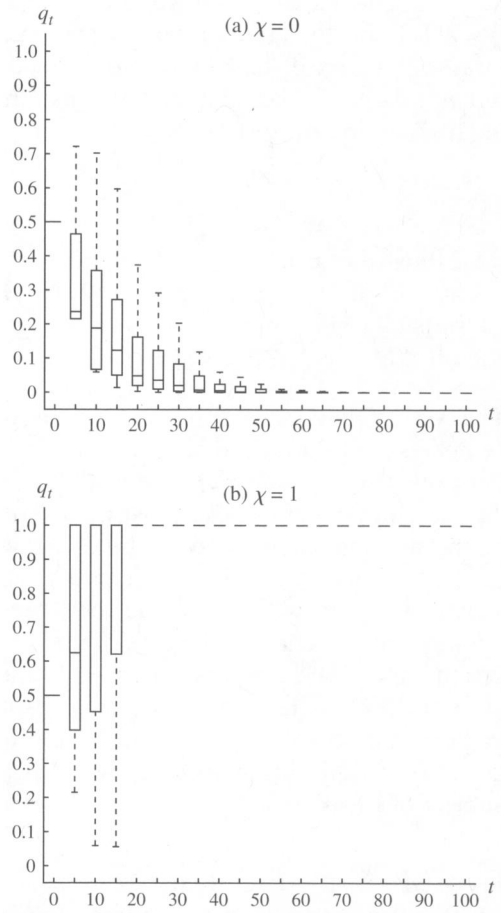
In summary, the linear and logit examples portrayed in Figure 1 are superficially similar, and yet the profit performance of the MBP is very bad in

Figure 2 Performance of the MBP in the Linear and Logit Examples



Note. In our linear example (11), the expected revenue loss $\Delta(T)$ increases linearly in the time horizon T, but in the logit example (12) it is bounded by a constant.

Figure 3 Evolution of the Belief Process of the MBP in the Logit Example



Notes. The box-and-whiskers plots in panels (a) and (b) show how the distribution of belief process q_t of the MBP evolves under hypotheses 0 and 1, respectively. At any given time period, the box displays the lower quartile, median, and upper quartile values for the 10,000 sample paths generated, and each whisker extends either to the most extreme data point or else for a distance equal to 1.5 times the inter-quartile range, whichever is smaller. Under either hypothesis, q_t rapidly finds the ambient demand model.

the linear example and is strikingly good in the logit example. These findings are explained by the theoretical development that follows.

4. Analysis of the Myopic Bayesian Policy

4.1. Incomplete Learning and Its Consequences

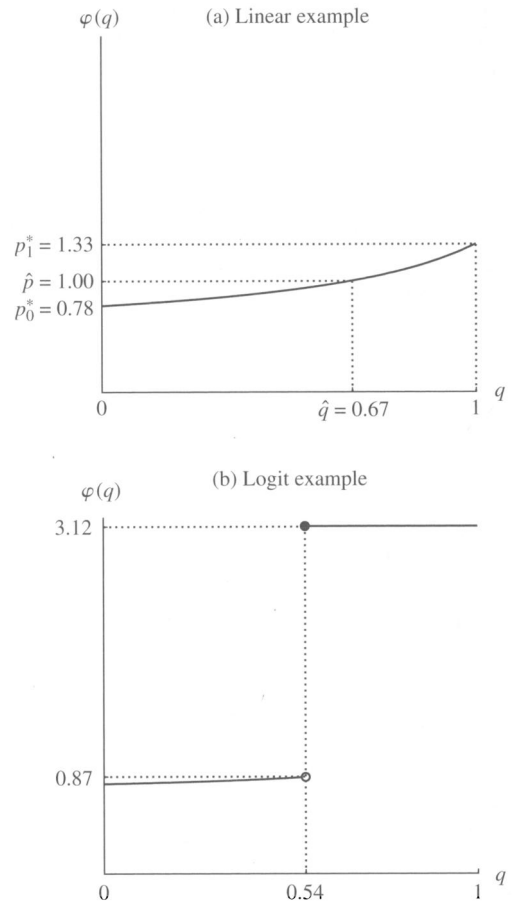
Let us start by identifying any belief $\hat{q} \in [0, 1]$ that makes the MBP charge the uninformative price $\hat{p}$ as a *confounding belief* for the MBP, with the following justification: If the seller enters period t with belief $q_t = \hat{q}$, then the MBP dictates price $p_t = \varphi(\hat{q}) = \hat{p}$, and because $\rho_0(\hat{p}) = \rho_1(\hat{p})$, formula (5) simply gives $q_{t+1} = q_t = \hat{q}$, and the same process repeats in every subsequent period; belief convergence is to $q_\infty = \hat{q}$. Thus, the seller never learns which demand hypothesis is true, and expected profit under the MBP is strictly suboptimal in every period. The following result proves that, if it exists, the confounding belief is unique.

Proposition 4 (Confounding Belief). *There exists at most one belief $\hat{q} \in [0, 1]$ such that $\varphi(\hat{q}) = \hat{p}$.*

We say that *learning is incomplete* under the MBP if $q_t \to \hat{q}$ as $t \to \infty$. If the incumbent price $\hat{p}$ lies between p_0^* and p_1^* (recall from §2 that $l < p_0^* < p_1^* < u$ by assumption), and if the myopic price function $\varphi(\cdot)$ is continuous, then existence of a confounding belief is guaranteed. In our linear example (11) the function $\varphi(\cdot)$ is continuous, as shown in Figure 4(a), and the confounding belief that corresponds to the uninformative price $\hat{p} = (a_0 - a_1)/(b_0 - b_1) = 1$ (see Figure 1(a)) is $\hat{q} = 0.67$.

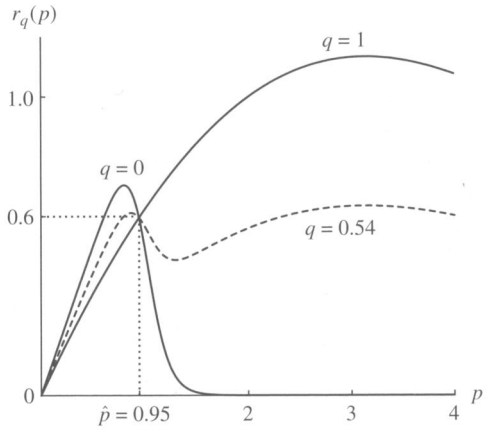
In contrast, *there does not exist* a confounding belief $\hat{q}$ for the logit example specified in (12) and pictured in Figure 1(b). This can be deduced from Figure 4(b), where the myopic price function $\varphi(q)$ jumps from 0.87 to 3.12 as q passes through the critical value of 0.54.

Figure 4 Myopic Price Function for the Linear and Logit Examples



Notes. In our linear example (11), the myopic price function $\varphi(\cdot)$: $[0, 1] \to [l, u]$ increases continuously, but in the logit example (12) it has a jump discontinuity; the discontinuity gap contains the uninformative price $\hat{p} = 0.95$, and hence there exists no confounding belief.

Figure 5 Single-Period Expected Profit Function for the Logit Example



Notes. The expected profit function $r_q(\cdot)$ is unimodal when q takes the extreme values of 0 and 1. It has, however, two global maxima when $q = 0.54$.

The unique uninformative price (that is, the incumbent price) for this example is $\hat{p} = 0.95$, and it lies in the interval of values over which $\varphi(\cdot)$ jumps, so the MBP will never charge prices within a neighborhood of $\hat{p}$.

The main reason for the absence of a confounding belief in our logit example (12) is the bimodality of the immediate revenue function. Figure 5 shows the single-period expected revenue function $r_q(\cdot)$ for three different values of q. For values of q near 0.5, that function is bimodal, with one peak near $p_0^* = 0.80$ and another one near $p_1^* = 3.13$. The first peak is larger for $q < 0.54$, the second one is larger for $q > 0.54$, and they are equal when $q = 0.54$. So the myopic price function $\varphi(\cdot)$ has a jump discontinuity at $q = 0.54$.

The obvious question to ask is the following: If a seller's prior belief is *different* from $\hat{q}$, is it possible that $q_t \to \hat{q}$ with positive probability under the MBP? Our next result shows that the answer is emphatically yes in the case of linear demand models because the seller's belief process $\{q_t\}$ cannot jump over the confounding belief $\hat{q}$, and subsequent discussion will show that the linear case is not exceptional.

Proposition 5 (Incomplete Learning Under Linear Demand). *Suppose that $\rho_0(\cdot)$ and $\rho_1(\cdot)$ are both linear and that a confounding belief $\hat{q}$ exists for the MBP. If $q_0 \le \hat{q}$ (respectively, $q_0 \ge \hat{q}$), then $q_t \le \hat{q}$ (respectively, $q_t \ge \hat{q}$) for all $t = 1, 2, \ldots$, where $\{q_t\}$ is the seller's belief process under the MBP.*

Returning to the simulation results presented in §3 for our linear example (11), we note that $\hat{q} > 0.5$ for that example, and hypothesis 1 is true in half of the simulation runs. Because q_t is a submartingale under the probability measure $\mathbb{P}_1^{\pi}(\cdot)$, it follows from Proposition 5 that $q_t \to \hat{q}$ almost surely when hypothesis 1 is true in the linear example. On the other hand, as will be explained later, even under hypothesis 0 one has that $q_t \to \hat{q}$ with positive probability.

To assess the loss due to incomplete learning, the following theorem shows that if a pricing policy π does *not* reveal almost surely which demand hypothesis is true, then its expected profit loss relative to a clairvoyant grows linearly in the time horizon. It will be shown later that this behavior can be "fixed" and the resulting variants of the MBP exhibit an expected profit loss that is bounded by a constant not depending on the horizon length T.

Theorem 1 (Profit Loss Due to Incomplete Learning). *Suppose that for a given policy π and a given hypothesis $i = 0, 1$, the limit belief q_∞ is neither almost surely 0 nor almost surely 1 and that $p_i^* \ne \hat{p}$. Then there exists a constant $\theta > 0$ such that $\Delta_i^{\pi}(T) \ge \theta T$ for all T.*

Proposition 5 actually allows one to calculate explicitly the distribution of the MBP limit belief q_∞ when both demand models are linear, as follows. For concreteness, assume that $q_0 \le \hat{q}$. Our starting point

is the following assertion: In light of Propositions 4 and 5, the only possible value for q_∞ under hypothesis 1 is $\hat{q}$, and the only possible values under hypothesis 0 are $\hat{q}$ and 0.

Next, observe that the posterior probabilities $\{q_t\}$ form a martingale under the probability measure $\mathbb{P}(\cdot) := q_0\mathbb{P}_1(\cdot) + (1-q_0)\mathbb{P}_0(\cdot)$. (Here we suppress the dependence of $\mathbb{P}$, $\mathbb{P}_0$, and $\mathbb{P}_1$ on the seller's pricing policy, which is taken to be the MBP throughout this discussion.) Denoting by $\mathbb{E}(\cdot)$ the expectation operator associated with $\mathbb{P}(\cdot)$, we then have from Doob's optional stopping theorem (Williams 1991, p. 100) that

$$\begin{aligned} q_0 = \mathbb{E}(q_\infty) &= q_0\mathbb{E}_1(q_\infty) + (1-q_0)\mathbb{E}_0(q_\infty) \\ &= q_0\hat{q} + (1-q_0)\hat{q}\mathbb{P}_0(q_\infty = \hat{q}). \end{aligned} \tag{13}$$

Solving this equation gives

$$\mathbb{P}_0(q_\infty = \hat{q}) = \frac{q_0(1-\hat{q})}{\hat{q}(1-q_0)}. \tag{14}$$

A virtually identical calculation gives the distribution of q_∞ when $q_0 \geq \hat{q}$, and readers will see that these calculations are equally valid for any other example where (a) a confounding belief $\hat{q}$ exists for the MBP and (b) the seller's belief process cannot jump over $\hat{q}$.

4.2. Extensions

The poor performance of the MBP is not limited to linear demand settings. The main intuition that can be carried over to more general settings is related to the ability of the belief process to "jump over" the confounding belief, which in turn depends on the extent to which the two possible demand models are "dissimilar." To measure dissimilarity between $\rho_0(\cdot)$ and $\rho_1(\cdot)$, we define the *rejection-likelihood-ratio function* $L(\cdot) := [1-\rho_0(\cdot)]/[1-\rho_1(\cdot)]$; this ratio measures the relative likelihood of a "no purchase" outcome under the two demand models. Using this we obtain the following generalization of Proposition 5.

Proposition 6 (Incomplete Learning Under Nonlinear Demand). *Suppose that the myopic price function $\varphi(\cdot)$ is continuously differentiable and that there exists a confounding belief $\hat{q}$ for the MBP. If $L(\varphi(q+h)) \leq L(\varphi(q))e^{4h}$ for all q and $q+h$ in $[0,1]$, and if $q_0 \leq \hat{q}$ (respectively, $q_0 \geq \hat{q}$), then $q_t \leq \hat{q}$ (respectively, $q_t \geq \hat{q}$) for all $t = 1, 2, \ldots$, where $\{q_t\}$ is the seller's belief process under the MBP.*

The key assumption of Proposition 6 requires that the dissimilarity metric $L(\cdot)$ increase slowly over its domain. To be precise, it defines a certain envelope such that if $L(\cdot)$ stays within that envelope, the belief process $\{q_t\}$ cannot jump over the confounding belief $\hat{q}$. Thus, if the demand models $\rho_0(\cdot)$ and $\rho_1(\cdot)$ do not become "too dissimilar" as we vary prices, the MBP is doomed to perform poorly. The opposite case is illustrated by the logit example (12), where the MBP performs well. There the myopic price function $\varphi(\cdot)$ has a jump discontinuity, which obviously violates the assumptions of Proposition 6. If both $\rho_0(\cdot)$ and $\rho_1(\cdot)$ come from the logit family of demand curves, it can be shown that the conditions of Proposition 6 are satisfied if and only if the potentially optimal prices p_0^* and p_1^* are sufficiently close to one another. Speaking more generally, in the numerical experiments that we have undertaken with demand models from various parametric families, incomplete learning under the MBP has proved to be the rule rather than the exception.

4.3. When the MBP Performs Well

The following result explains the good performance of the MBP in our logit example; see Figures 2(b) and 3 in §3.

Theorem 2 (Profit Loss Bounded by a Constant). *If there does not exist a belief $q \in [0,1]$ such that $\hat{p} \in \arg\max\{r_q(p), l \leq p \leq u\}$, then there exists a finite positive constant C such that $\Delta_i^\pi(\infty) \leq C$ for $i = 0, 1$, where $\Delta_i^\pi(\cdot)$ is given in (8), for the pricing policy $\pi = MBP$.*

Remark. The hypothesis of the above theorem is essentially that there exists no confounding belief $\hat{q}$. To be more precise, it requires that the uninformative price $\hat{p}$ not be myopically optimal for any belief $q \in [0,1]$, even in the case of bimodal expected revenue functions.

5. Discriminative Policies and Their Performance

A policy π is said to be *δ-discriminative* if

$$|\rho_0(\pi_t(q)) - \rho_1(\pi_t(q))| > \delta \quad \text{for all } t = 1, 2, \ldots \text{ and } 0 \leq q \leq 1 \tag{15}$$

and to be *discriminative* if it is δ-discriminative for some $\delta > 0$. The following proposition is used in the proof of Theorem 2. It will also guide our search for a modified version of the MBP that achieves the same performance guarantee announced in Theorem 2 (namely, that the expected revenue loss is bounded by a constant), but without any restrictions on the demand models.

Proposition 7 (Rate of Learning). *If π is a discriminative policy, then there exist constants μ, $\lambda > 0$ such that*

$$\mathbb{E}_0^\pi(q_t) \leq \mu\exp(-\lambda t) \quad \text{and} \quad \mathbb{E}_1^\pi(1-q_t) \leq \mu\exp(-\lambda t) \quad \text{for all } t = 1, 2, \ldots. \tag{16}$$

To see how Proposition 7 applies to Theorem 2, one need only note the following: If the uninformative price $\hat{p}$ is not optimal for any $q \in [0,1]$, then the MBP is a discriminative policy.

Table 1 $\Delta(T)$ **Under CMBP(ϵ) for Various Values of ϵ and T**

ϵ	$T=10$	$T=100$	$T=1{,}000$	$T=2{,}000$	$T=3{,}000$	$T=5{,}000$	$T=10{,}000$
0.05	0.6	6.0	32.3	37.9	39.5	39.5	39.6
0.10	0.6	5.7	17.3	17.8	17.7	18.3	18.3
0.15	0.7	5.2	10.5	10.5	10.4	10.4	10.5
0.20	0.8	4.6	7.2	7.0	6.9	7.0	6.9

Notes. If the constraint parameter ϵ is large enough to include one of the potentially optimal prices p_i^* within the forbidden price interval $[\hat{p}-\epsilon, \hat{p}+\epsilon]$, then the performance metric $\Delta(T)$ will grow linearly in T. Thus, for any set of model parameters, there exists a natural upper bound on ϵ such that CMBP(ϵ) cannot perform well if ϵ exceeds that bound. The value of this upper bound in this example is 0.22.

Remark. The content of Proposition 7 is reminiscent of results in the hypothesis testing literature. To wit, when there are a finite number of hypotheses, as in our setup, and when these hypotheses are "well separated," which is the case under a discriminative policy, then in some generality one expects the error probability to decay exponentially fast in the number of observations (see Dembo and Zeitouni 1998, §3.4). Indeed, our proof of Proposition 7 uses an information theoretic measure of "distance" between the alternative demand models, and this measure of separation plays a central role in the hypothesis testing literature. A key feature that distinguishes our problem from more traditional hypothesis testing formulations is the presence of dependence in the sequence of observations. Of course, any sensible policy for learning-and-earning will give rise to such dependence, and it is handled in the proof of Proposition 7 by means of a martingale large deviation bound (the Azuma–Hoeffding inequality; Williams 1991, p. 237).

6. A Constrained Variant of the MBP (CMBP)

Our first approach to avoiding incomplete learning follows in a straightforward manner from the analysis of the MBP in §4. As suggested by the logit example, one can simply preclude prices that are in a neighborhood of the uninformative one. Given a policy parameter $\epsilon > 0$, let

$$\hat{\varphi}(q) := \sup\{\operatorname{argmax}\{r_q(p)\colon l \le p \le u \text{ and } |p-\hat{p}| \ge \epsilon\}\} \tag{17}$$

for $0 \le q \le 1$. The pricing policy π that has $\pi_t(\cdot) = \hat{\varphi}(\cdot)$ for all $t = 1, 2, \ldots$ is called the *constrained* variant of the MBP, with associated *constraint parameter* ϵ, hereafter abbreviated CMBP(ϵ). The following theorem states that the profit loss under CMBP(ϵ), relative to the profit earned by a clairvoyant, is bounded by a constant. This follows in a fairly straightforward manner because the constrained version of the MBP is a discriminative policy.

Theorem 3 (Performance of Constrained MBP). *Assume that $p_i^* \ne \hat{p}$ for $i = 0, 1$ and that $\epsilon > 0$ is small enough to ensure that the argument set in* (17) *contains p_0^* and p_1^*. Then there exists a finite positive constant C such that $\Delta_i^\pi(\infty) \le C$ for $i = 0, 1$, where $\Delta_i^\pi(\cdot)$ is given in* (8)*, for the pricing policy $\pi = CMBP(\epsilon)$.*

Table 1 illustrates the performance of CMBP(ϵ) in the linear example (11). To be more specific, Table 1 gives simulation estimates for the metric $\Delta(\cdot) = \frac{1}{2}\Delta_0^\pi(\cdot) + \frac{1}{2}\Delta_1^\pi(\cdot)$ under CMBP(ϵ); all of these simulation estimates are based on 100,000 independent replications of the relevant sales sequence. We tabulate $\Delta(T)$ for different values of the constraint parameter ϵ and different horizon lengths T. The performance $\Delta(\cdot)$ is essentially flat beyond $T = 2{,}000$ periods, and the minimum value of $\Delta(T)$ for large T is approximately 7, achieved by taking $\epsilon = 0.2$.

This successful "fix" for the MBP depends crucially on the assumption that the optimal price under each demand hypothesis is distinct from the uninformative price $\hat{p}$. We now show that if the optimal price under either demand hypothesis *does* equal $\hat{p}$, then the performance associated with CMBP in Theorem 3 (that is, expected revenue loss bounded by a constant) cannot be achieved by *any* pricing policy.

Proposition 8. *Assume that a confounding belief $\hat{q}$ exists and that $p_i^* = \hat{p}$ for some $i = 0, 1$. Then $\Delta_i^\pi(\infty) = \infty$ for any pricing policy π and some $i = 0, 1$.*

The situation treated in Proposition 8, where $p_i^* = \hat{p}$ for either $i = 0$ or $i = 1$, can be viewed as a particular instance of the following. Suppose that the seller's prior distribution over possible demand models is general, not necessarily binary, and for every open interval I containing $\hat{p}$ it assigns positive probability to models whose associated optimal price lies within I. Because a CMBP policy forbids prices within a fixed interval surrounding $\hat{p}$, its associated expected loss must grow linearly under any such prior distribution. A desirable generalization of the CMBP is therefore to shrink the excluded interval as more information about the ambient demand model accumulates. This is the subject of the next section.

7. A Generalized Version of CMBP

We consider in this section a pricing policy that involves a sequence of constraint parameters $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots)$. The corresponding policy will be referred

to as the *flexibly constrained* variant of the MBP, abbreviated FMBP($\boldsymbol{\epsilon}$). For a definition of the policies, first let

$$\hat{\varphi}_t(q) := \sup \arg\max\{r_q(p): l \le p \le u \text{ and } |p-\hat{p}| \ge \epsilon_t\} \tag{18}$$

for $t = 1, 2, \ldots$ and $0 \le q \le 1$. Then FMBP($\boldsymbol{\epsilon}$) is the pricing policy π that has $\pi_t(\cdot) = \hat{\varphi}_t(\cdot)$ for all $t = 1, 2, \ldots$. The following analog of Theorem 3 proves that if the constraint parameters are suitably chosen, then the profit loss of FMBP is also bounded by a constant.

THEOREM 4 (PERFORMANCE OF FLEXIBLY CONSTRAINED MBP). *Assume that $p_i^* \neq \hat{p}$ for $i = 0, 1$. Construct the sequence $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots)$ by setting $\epsilon_t = \epsilon_1 t^{-\alpha}$ for any fixed $\alpha \in (0, \frac{1}{4})$ and $\epsilon_1 > 0$ small enough to ensure that the argument set in* (18) *contains p_0^* and p_1^* for all t. Then there exists a finite positive constant C such that $\Delta_i^{\pi}(\infty) \le C$ for $i = 0, 1$, where $\Delta_i^{\pi}(\cdot)$ is given in* (8)*, for the pricing policy $\pi =$ FMBP($\boldsymbol{\epsilon}$).*

REMARK. The constant C is explicitly identified in the proof and is given in terms of the problem primitives and the constraint parameters $\boldsymbol{\epsilon}$.

One can view FMBP as augmenting the MBP with price experiments in the form of minimum tolerable deviations from the incumbent price $\hat{p}$. Whenever the difference between the MBP price and the incumbent price falls below the minimum tolerable deviation, FMBP exercises a test price. Because $\epsilon_t \to 0$ as $t \to \infty$, and because $\hat{p}$ is not optimal under either demand hypothesis (this is crucial), the FMBP price and the MBP price eventually coincide. The constraint parameters specified in Theorem 4 decrease slowly enough to avoid the incomplete learning trap but still allow the expected revenue loss to be bounded by a constant.

Table 2 displays the simulated performance of FMBP($\boldsymbol{\epsilon}$) in the linear example. Because the CMBP performance did not change beyond 2,000 periods, we simply fix the horizon length at $T = 2{,}000$, tabulating $\Delta(2{,}000)$ for different constraint sequences $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots)$. Varying the values of ϵ_1 and α, we observe that the best performance is achieved at the parameter values $\epsilon_1 = 0.2$ and $\alpha = 0.0025$. Under FMBP($\boldsymbol{\epsilon}$) with these tuning parameters, the total expected loss due to initial model uncertainty equals the expected profit from just seven to eight sales opportunities. In the context of retail financial services, where even a small provider has many hundreds of sales opportunities per day, this can reasonably be called a negligible loss.

Table 2 $\Delta(2{,}000)$ **Under FMBP($\boldsymbol{\epsilon}$) for Various Values of ϵ_1 and α**

ϵ_1	$\alpha = 0.0025$	$\alpha = 0.0625$	$\alpha = 0.125$	$\alpha = 0.1875$	$\alpha = 0.2475$
0.05	38.94	52.41	66.34	78.06	86.62
0.10	17.98	26.45	39.02	53.22	67.23
0.15	10.67	15.36	23.84	36.09	49.93
0.20	7.20	10.17	15.60	24.21	36.67

8. Summary of Main Contributions

The three main contributions of this paper are (a) its elucidation of incomplete learning, specifically as a consequence of myopic pricing under model uncertainty; (b) its development of CMBP policies as a means of "fixing" the myopic policy in a binary-prior setting; and (c) its investigation of FMBP policies, whose added flexibility makes them potentially attractive in a general setting where the seller has an arbitrary prior distribution over demand models.

To the best of our knowledge, the results in this paper are the first about incomplete learning due to myopic pricing. However, there is a substantial literature on negative effects of myopic decision making in a more general context, including negative effects that can reasonably be described as incomplete learning. Representative of that literature is the early paper by Lai and Robbins (1982) on "iterated least squares," which was stimulated by the numerical investigation of Anderson and Taylor (1976). The general problem addressed by Lai and Robbins (1982), and in other similar work, does not bear upon dynamic pricing in an obvious or direct way. In this paper we have exploited the special structure of our stylized model, featuring a binary response to each offered price and a Bayesian formulation with a binary-prior distribution, to prove sharp results on incomplete learning under myopic or greedy pricing. Those results are presented in §4, where, in particular, Proposition 5 shows that incomplete learning is ubiquitous when a linear demand model is assumed (the case that has dominated in previous research).

With regard to (b), our most notable result (Theorem 3 in §6) is that the expected revenue loss under a simple CMBP policy is bounded by a constant, independent of the horizon length, assuming that the incumbent price is not optimal under either of the seller's two demand hypotheses. Asymptotic performance bounds are a staple in the theory of online decision making under model uncertainty, which has branches in statistics, machine learning, and adaptive control; in that literature the performance measure we use in this paper is usually referred to as *regret*. However, constant bounds on the regret (as opposed to bounds that grow as a fractional power of the horizon length or grow as the logarithm of the horizon length) are rare, and we do not know of any bound comparable to ours in the literature of dynamic pricing. If the qualifying assumption of Theorem 3 is removed (that is, if the optimal price under one demand hypothesis *does* equal the incumbent price), we have also

shown that *no policy* can achieve expected revenue losses bounded by a constant (this is Proposition 8 in §6), which suggests that the strong conclusion of Theorem 3 cannot hold with a general prior distribution over demand models.

With regard to (c), our central result (Theorem 4 of §7) is an extension of Theorem 3. It shows that if the constraint parameters of an FMBP converge to zero at a properly chosen rate, then one can still achieve expected revenue loss bounded by a constant in our binary-prior setting. The FMBP constraints on successive prices specify what might be called a minimal degree of price experimentation, relative to the incumbent price. With a general or arbitrary prior distribution over demand models, as opposed to the binary prior assumed in this paper, we conjecture that the best achievable asymptotic loss rate (whatever that may be) can be gotten with a policy of the FMBP form.

Acknowledgments

The authors are indebted to Robert Phillips, chief science officer of Nomis Solutions, for suggesting the topic explored in this paper and for valuable feedback during the course of their research. Research was partially supported by the National Science Foundation [Grant DMI-0447562].

Appendix A. More on Discriminative Policies

In this section we generalize the definition of discriminative policies and prove an important result regarding the learning rate. A price $p \in [l, u]$ is said to be *δ-discriminative* if

$$|\rho_0(p) - \rho_1(p)| > \delta, \tag{A1}$$

and given a real number sequence $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots)$, a policy π is said to be *$\boldsymbol{\delta}$-discriminative* if the price $\pi_t(q)$ is δ_t-discriminative for all $t = 1, 2, \ldots$ and $q \in [0, 1]$. The following lemma provides the key ingredient for establishing the main results in §§5–7.

Lemma A.1. *Fix $\alpha \in [0, 1/4)$ and $\delta > 0$, and then define a sequence $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots)$ by setting $\delta_t = \delta t^{-\alpha}$ for all $t = 1, 2, \ldots$. Then, for any $\boldsymbol{\delta}$-discriminative policy π, there exist constants $\mu, \lambda > 0$ such that*

$$\mathbb{E}_0^\pi(q_t) \le \mu \exp(-\lambda t^{1-4\alpha}) \quad \text{and} \quad \mathbb{E}_1^\pi(1 - q_t) \le \mu \exp(-\lambda t^{1-4\alpha}) \quad \text{for all } t = 1, 2, \ldots. \tag{A2}$$

Proof. Assume without loss of generality that $\chi = 0$. (The analysis that follows can be repeated verbatim for the case $\chi = 1$.) The proof of all auxiliary results stated below is deferred to Appendix E, and the following abbreviated notation will be used repeatedly: Put $\rho_{i,t} := \rho_i[\pi_t(q_{t-1})]$ for all $i = 0, 1$, and $t \in \mathbb{N}$, and $\bar{y} := 1 - y$ for all $y \in [0, 1]$. To simplify notation further, we also put $\alpha_k := \rho_{1,k}/\rho_{0,k}$ and $\beta_k := (1 - \rho_{0,k})/(1 - \rho_{1,k})$ for all $k \in \mathbb{N}$.

We will make use of the following expression for the belief q_t (whose proof is straightforward and hence omitted). For all $t = 1, 2, \ldots$

$$q_t = \sum_{x \in \{0,1\}^t} \frac{q_0 A_t(x)}{q_0 A_t(x) + (1 - q_0) B_t(x)} I\{(X_1, \ldots, X_t) = x\}, \tag{A3}$$

where $A_t(x) := \prod_{k=1}^t \alpha_k^{x_k}$ and $B_t(x) := \prod_{k=1}^t \beta_k^{1-x_k}$.

Fix $q_0 \in (0, 1)$ and fix a $\boldsymbol{\delta}$-discriminative policy π such that $\delta_t = \delta_1 t^{-\alpha}$ for all t. Taking an expectation of the above expression yields

$$\begin{aligned}
&\mathbb{E}_0^\pi[q_t] \\
&= \sum_{x \in \{0,1\}^t} \frac{q_0 A_t(x)}{q_0 A_t(x) + (1 - q_0) B_t(x)} \mathbb{P}_0^\pi[(X_1, \ldots, X_t) = x] \\
&= \mathbb{E}_0^\pi\left[\frac{q_0 \prod_{k=1}^t \alpha_k^{X_k}}{q_0 \prod_{k=1}^t \alpha_k^{X_k} + \bar{q}_0 \prod_{k=1}^t \beta_k^{1-X_k}}\right] \\
&= \mathbb{E}_0^\pi\left[\frac{1}{1 + (\bar{q}_0/q_0) \prod_{k=1}^t \beta_k^{1-X_k} \alpha_k^{-X_k}}\right] \\
&= \mathbb{E}_0^\pi\left[\frac{1}{1 + (\bar{q}_0/q_0) \exp(\sum_{k=1}^t [(1 - X_k) \log \beta_k - X_k \log \alpha_k])}\right].
\end{aligned} \tag{A4}$$

Let L_t denote the argument of the $\exp(\cdot)$ in the denominator in (A4). Simple algebra, and using the definition of α_k, β_k above, then yields

$$\begin{aligned}
L_t &= \sum_{k=1}^t [(1 - X_k) \log \beta_k - X_k \log \alpha_k] \\
&= \sum_{k=1}^t [(1 - X_k - \bar{\rho}_{0,k}) \log \beta_k - (X_k - \rho_{0,k}) \log \alpha_k] \\
&\quad + \sum_{k=1}^t [\bar{\rho}_{0,k} \log \beta_k - \rho_{0,k} \log \alpha_k] \\
&= \sum_{k=1}^t \left[(\rho_{0,k} - X_k) \log \frac{\bar{\rho}_{0,k}}{\bar{\rho}_{1,k}} + (X_{k+1} - \rho_{0,k}) \log \frac{\rho_{0,k}}{\rho_{1,k}}\right] \\
&\quad + \sum_{k=1}^t \left[\bar{\rho}_{0,k} \log \frac{\bar{\rho}_{0,k}}{\bar{\rho}_{1,k}} + \rho_{0,k} \log \frac{\rho_{0,k}}{\rho_{1,k}}\right] \\
&= \sum_{k=1}^t \left[(X_k - \rho_{0,k}) \log \frac{\rho_{0,k} \bar{\rho}_{1,k}}{\bar{\rho}_{0,k} \rho_{1,k}}\right] \\
&\quad + \sum_{k=1}^t \left[\bar{\rho}_{0,k} \log \frac{\bar{\rho}_{0,k}}{\bar{\rho}_{1,k}} + \rho_{0,k} \log \frac{\rho_{0,k}}{\rho_{1,k}}\right].
\end{aligned}$$

The following lemma allows us to bound L_t from below.

Lemma A.2. *For any $\boldsymbol{\delta}$-discriminative policy π,*

$$\bar{\rho}_{0,k} \log \frac{\bar{\rho}_{0,k}}{\bar{\rho}_{1,k}} + \rho_{0,k} \log \frac{\rho_{0,k}}{\rho_{1,k}} \ge 2\delta_k^2 \quad \text{for all } k = 1, 2, \ldots.$$

Combining the expression derived above for L_t with this inequality and recalling that the sequence $\boldsymbol{\delta}$ is nonincreasing, we have the following:

$$\begin{aligned}
L_t &\ge \sum_{k=1}^t \left[(X_k - \rho_{0,k}) \log \frac{\rho_{0,k} \bar{\rho}_{1,k}}{\bar{\rho}_{0,k} \rho_{1,k}}\right] + 2 \sum_{k=1}^t \delta_k^2 \\
&\ge \sum_{k=1}^t \left[(X_k - \rho_{0,k}) \log \frac{\rho_{0,k} \bar{\rho}_{1,k}}{\bar{\rho}_{0,k} \rho_{1,k}}\right] + 2t\delta_t^2.
\end{aligned}$$

We now turn to analyzing the sum on the right-hand side above. To that end, put $M_0 = 0$ and

$$M_t := \sum_{k=1}^t \left[(X_k - \rho_{0,k}) \log \frac{\rho_{0,k} \bar{\rho}_{1,k}}{\bar{\rho}_{0,k} \rho_{1,k}}\right] \quad \text{for all } t = 1, 2, \ldots.$$

Fix $\xi > 0$ and consider the event $\mathcal{J}_{t,\xi} := \{|M_t| < \xi t\}$. Using the above and recalling (A4), we deduce that

$$\begin{aligned}\mathbb{E}_0^\pi[q_t] &= \mathbb{E}_0^\pi\left[\frac{1}{1+(\bar{q}_0/q_0)\exp(L_t)}\right]\\ &\le \mathbb{E}_0^\pi\left[\frac{1}{1+(\bar{q}_0/q_0)\exp(M_t+2t\delta_t^2)}; \mathcal{J}_{t,\xi}\right]\\ &\quad + \mathbb{E}_0^\pi\left[\frac{1}{1+(\bar{q}_0/q_0)\exp(M_t+2t\delta_t^2)}; \mathcal{J}_{t,\xi}^c\right]\\ &\le \frac{1}{1+(\bar{q}_0/q_0)\exp(-\xi t+2t\delta_t^2)} + \mathbb{P}_0^\pi(\mathcal{J}_{t,\xi}^c),\end{aligned}$$

where the second inequality follows from the definition of the event $\mathcal{J}_{t,\xi}$. To finish the proof, we need a bound on the probability of the complement of this event.

Lemma A.3. *For any policy π, there exists a positive real number γ such that*

$$\mathbb{P}_0^\pi(\mathcal{J}_{t,\xi}^c) \le 2\exp\left(-\frac{1}{2\gamma}\xi^2 t\right) \quad \text{for all } t = 1, 2, \ldots.$$

Combining this with the above, we have for all t the following:

$$\begin{aligned}\mathbb{E}_0^\pi[q_t] &\le \frac{1}{1+(\bar{q}_0/q_0)\exp(-t\xi+2t\delta_t^2)} + 2\exp\left[-\frac{1}{2\gamma}\xi^2 t\right]\\ &\le \frac{q_0}{\bar{q}_0}\exp(t\xi - 2t\delta_t^2) + 2\exp\left[-\frac{1}{2\gamma}\xi^2 t\right]\\ &\le \mu e^{-\lambda_t(\xi)t},\end{aligned}$$

where $\mu = 2\max\{2, q_0/\bar{q}_0\}$ and $\lambda_t(\xi) = \min\{2\delta_t^2 - \xi, \xi^2/2\gamma\}$. Setting $\xi = \delta_t^2 = (\delta t^{-\alpha})^2 > 0$, we have $t^{4\alpha}\lambda_t(\xi) = \delta^2 \min\{t^{2\alpha}, \delta^2/2\gamma\} \ge \delta^2 \min\{1, \delta^2/2\gamma\}$. Choosing $\lambda \equiv \delta^2 \min\{1, \delta^2/2\gamma\}$ completes the proof. □

Appendix B. Proofs of Propositions 1–3

Proof of Proposition 1. First, recall that the incumbent price $\hat{p}$ satisfies $\rho_0(\hat{p}) = \rho_1(\hat{p})$. Now, let $\tilde{p}$ be an arbitrary price in $[p_0^*, p_1^*]$ such that $\rho_1(\tilde{p}) = \rho_0(\tilde{p})$. Because $\varepsilon_1(p) < 1 < \varepsilon_0(p)$ for all $p \in [p_0^*, p_1^*]$, we deduce that $\rho_1'(\tilde{p}) - \rho_0'(\tilde{p}) > 0$. Thus, the function $p \mapsto \rho_1(p) - \rho_0(p)$ is locally increasing around $\tilde{p}$ and vanishes at that point. This implies that there can be at most one price $p \in [p_0^*, p_1^*]$ satisfying $\rho_1(p) = \rho_0(p)$. Because the incumbent price $\hat{p}$ already satisfies that condition, we conclude that $\hat{p}$ is the unique price $p \in [p_0^*, p_1^*]$ such that $\rho_1(p) = \rho_0(p)$. □

Proof of Proposition 2. Assume without loss of generality that $\chi = 0$. Let $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$ and note that q_t is bounded and hence $\mathcal{F}_t$-measurable under both $\mathbb{P}_0^\pi$ and $\mathbb{P}_1^\pi$. Moreover, putting $\rho_{i,t} = \rho_i[\pi_t(q_{t-1})]$ for all $i = 0, 1$, $t \in \mathbb{N}$, and $\bar{y} = 1 - y$ for all $y \in [0, 1]$, one has

$$\begin{aligned}\mathbb{E}_0^\pi[q_{t+1} \mid \mathcal{F}_t] &= \rho_{0,t}\cdot\frac{q_t\rho_{1,t}}{q_t\rho_{1,t}+\bar{q}_t\rho_{0,t}} + \bar{\rho}_{0,t}\cdot\frac{q_t\bar{\rho}_{1,t}}{q_t\bar{\rho}_{1,t}+\bar{q}_t\bar{\rho}_{0,t}}\\ &= q_t - \frac{q_t^2\bar{q}_t(\rho_{1,t}-\rho_{0,t})^2}{(q_t\rho_{1,t}+\bar{q}_t\rho_{0,t})(q_t\bar{\rho}_{1,t}+\bar{q}_t\bar{\rho}_{0,t})}\end{aligned}$$

for all t. Thus, $\{q_t\}$ is a bounded supermartingale under $\mathbb{P}_0^\pi$. A standard result in martingale theory implies that $\{q_t\}$ converges almost surely to a finite random variable q_∞ (see Williams 1991, p. 109). □

Proof of Proposition 3. Because $\varepsilon_0(\cdot)$ and $\varepsilon_1(\cdot)$ are strictly increasing, $r_0(\cdot)$ and $r_1(\cdot)$ are strictly quasi-concave. Therefore, $r_q(\cdot)$ for any $q \in [0, 1]$ is increasing on $[l, p_0^*) \cap [l, p_1^*) = [l, p_0^*)$ and decreasing on $(p_0^*, u] \cap (p_1^*, u] = (p_1^*, u]$. Hence, for any $q \in [0, 1]$ we have $\varphi(q) \in [p_0^*, p_1^*]$. In other words, $\varphi([0, 1]) \subseteq [p_0^*, p_1^*]$. Now, by the definition of price elasticity we have for all $p \in [p_0^*, p_1^*]$ and $i \in \{0, 1\}$ that $1 - \varepsilon_i(p)$ has the same sign as $r_i'(p)$. Because $\varepsilon_i(p_i^*) = 1$ for $i \in \{0, 1\}$ and $\varepsilon_0(\cdot)$ and $\varepsilon_1(\cdot)$ are strictly increasing, we conclude that $r_1'(p)$ is positive for all $p \in [p_0^*, p_1^*]$ whereas $r_0'(p)$ is negative over the same price range. Recall that $\varphi(q) \in \arg\max\{r_q(p)\colon l \le p \le u\}$ for all $q \in [0, 1]$ and note that

$$\frac{\partial^2 r_q(p)}{\partial p \partial q} = r_1'(p) - r_0'(p).$$

Because $r_1'(p) > 0$ and $r_0'(p) < 0$ for all $p \in [p_0^*, p_1^*]$, one has $\partial^2 r_q(p)/\partial p \partial q > 0$ for all $p \in [p_0^*, p_1^*]$; that is $r_q(p)$ is supermodular in (p, q) on $[p_0^*, p_1^*] \times [0, 1]$. Hence, by the Topkis (1978) theorem, we deduce that $\varphi(q)$ is nondecreasing in q. □

Appendix C. Proofs of Propositions 4–6 and Theorem 1

Proof of Proposition 4. Note that if $\hat{p} \in [p_0^*, p_1^*]$, then $\hat{p} = \varphi(\hat{q}) = \sup\arg\max\{r_{\hat{q}}(p)\colon l \le p \le u\}$. Thus, by the first-order conditions of optimality,

$$\hat{q} = -\frac{\rho_0(\hat{p}) + \hat{p}\rho_0'(\hat{p})}{\hat{p}[\rho_1'(\hat{p}) - \rho_0'(\hat{p})]}. \tag{C1}$$

On the other hand, if $\hat{p} \notin [p_0^*, p_1^*]$, then there does not exist a confounding belief $\hat{q}$. This completes the proof. □

Proof of Proposition 5. We have the uniqueness of the confounding belief $\hat{q}$ by Proposition 4. Moreover, as argued in the proof of Proposition 1, the function $p \mapsto \rho_1(p) - \rho_0(p)$ is locally increasing around $\hat{p}$ and vanishes at $\hat{p}$. Because $p = \hat{p}$ is the unique price such that $\rho_1(p) = \rho_0(p)$, we have

$$\begin{aligned}\rho_0(p) > \rho_1(p) \quad \forall p < \hat{p} \quad \text{and}\\ \rho_0(p) < \rho_1(p) \quad \forall p > \hat{p}.\end{aligned} \tag{C2}$$

Without loss of generality, assume that the seller's belief is $q_t = q \in [0, \hat{q})$ at the end of a given period t and that $p_{t+1} = \varphi(q) = p$. We will show that q_{t+1} cannot exceed $\hat{q}$.

Case 1: $X_{t+1} = 1$. We know by Proposition 3 that the myopic price $\varphi(q)$ is monotone increasing in q. Therefore, $p = \varphi(q) \le \hat{p}$ because $q \le \hat{q}$. Now, by plugging $X_{t+1} = 1$ into (5), we have q_{t+1} expressed in terms of $q_t = q$ and $p_{t+1} = p$,

$$q_{t+1}(q, p) = \frac{q\rho_1(p)}{q\rho_1(p) + (1-q)\rho_0(p)}.$$

Because $p \le \hat{p}$, by (C2) we have that $(1 - q)\rho_1(p) < (1 - q)\rho_0(p)$. Thus, $q_{t+1}(q, p) < q \le \hat{q}$.

Case 2: $X_{t+1} = 0$. We will show that $q_{t+1}(q, p) \le \hat{q}$; i.e.,

$$q_{t+1}(q, p) = \frac{q\bar{\rho}_1(p)}{q\bar{\rho}_1(p) + (1-q)\bar{\rho}_0(p)} \le \hat{q},$$

or equivalently,

$$\hat{q}(1-q)\bar{\rho}_0(p) - q(1-\hat{q})\bar{\rho}_1(p) \ge 0. \tag{C3}$$

In the case of of a linear demand model, the uninformative price and the myopic price can be explicitly expressed as $\hat{p} = (a_0 - a_1)/(b_0 - b_1)$ and $p = \varphi(q) = a_q/(2b_q)$, where for brevity we set $a_q := qa_1 + (1-q)a_0$ and $b_q := qb_1 + (1-q)b_0$. Furthermore, we get by (C1) that

$$\hat{q} = \frac{a_0 b_0 - 2a_1 b_0 + a_0 b_1}{(b_0 - b_1)(a_0 - a_1)}.$$

Now, we plug the above expressions of $p = \varphi(q)$ and $\hat{q}$ into the left-hand side of (C3) and deduce by simple algebra that (C3) is equivalent to

$$[\hat{q}(1-q) - q(1-\hat{q})] \cdot [qb_1(2-a_1) + (1-q)b_0(2-a_0)] \geq 0.$$

To show that the above statement is correct, we recall that p_0^* and p_1^* are interior points of the interval $[l, u]$. We also know that $\rho_0(p_0^*) \leq 1$ and $\bar{\rho}_1(p_1^*) \leq 1$, which imply that $a_0 \leq 2$ and $a_1 \leq 2$. As a result, the left-hand side of the above expression is indeed nonnegative for $q < \hat{q}$, and hence the MBP belief process q_t cannot "jump over" the confounding belief $\hat{q}$.

The above argument can be applied verbatim to the case where $q \in (\hat{q}, 1]$. This completes the proof. □

Proof of Theorem 1. We begin by showing that the given policy π cannot charge a δ-discriminative price infinitely often. Define the event $\mathcal{D}_{t,\delta}^{\pi} := \{|\rho_1(p_t) - \rho_0(p_t)| > \delta\}$ for all $t = 1, 2, \ldots$ and $\delta > 0$. Assume toward a contradiction that with probability one $\mathcal{D}_{t,\delta}^{\pi}$ occurs infinitely often. This implies that there exists a subsequence $q_{t(k)}$ of q_t such that $|\rho_1(p_{t(k)}) - \rho_0(p_{t(k)})| > \delta$ for all $k \in \mathbb{N}$. Thus, a straightforward generalization of Lemma A.2 and direct application of Lemma A.3 lead to

$$\mathbb{E}_0^{\pi}[q_{t(k)}] \leq \frac{1}{1 + (\bar{q}_0/q_0)\exp(-t(k)\epsilon + 2k\delta^2)} + 2\exp\left[-\frac{1}{2\gamma}\epsilon^2 t(k)\right].$$

As argued in the proof of Lemma A.1, we conclude that there exist constants $\mu, \lambda > 0$ such that $\mathbb{E}_0^{\pi}[q_{t(k)}] \leq \mu e^{-\lambda k}$. Hence, $q_\infty = 0$ almost surely, which contradicts the fact that q_∞ is neither 0 nor 1, almost surely. Therefore, with positive probability $\mathcal{D}_{t,\delta}^{\pi}$ occurs only finitely often. Denoting the number of occurrences in T periods by $D_T := \sum_{t=1}^{T} I\{\mathcal{D}_{t,\delta}^{\pi}\}$, we have that $\mathbb{P}_0^{\pi}(D_\infty < \infty) > 0$.

Given $\delta > 0$, we know by continuity of $\rho_0(\cdot)$ and $\rho_1(\cdot)$ that there exist $\tilde{\delta} > 0$ such that $|p_t - \hat{p}| < \tilde{\delta}$ in any period t where $\mathcal{D}_{t,\delta}^{\pi}$ does not occur. Here we let $\hat{r}_i := \max\{r_i(p), \hat{p} - \tilde{\delta} \leq p \leq \hat{p} + \tilde{\delta}\}$. Because $p_0^* \neq \hat{p}$, we can choose δ sufficiently small so that $\hat{r}_0 < r_0(p_0^*)$. Then, by Fatou's lemma, we have

$$\begin{aligned}\liminf_{T\to\infty} \frac{\Delta_0^{\pi}(T)}{T} &\geq \mathbb{E}_0^{\pi}\left[\liminf_{T\to\infty}\left(1 - \frac{\sum_{t=1}^{T} r_0(p_t)}{T r_0(p_0^*)}\right)\right] \\ &\geq \mathbb{E}_0^{\pi}\left[\liminf_{T\to\infty}\left(1 - \frac{D_T r_0(p_0^*) + (T - D_T)\hat{r}_0}{T r_0(p_0^*)}\right)\right] \\ &\geq \left(1 - \frac{\hat{r}_0}{r_0(p_0^*)}\right)\mathbb{P}_0^{\pi}(D_\infty < \infty). \qquad \text{(C4)}\end{aligned}$$

The proof is complete by setting $\theta := (1 - \hat{r}_0/r_0(p_0^*)) \cdot \mathbb{P}_0^{\pi}(D_\infty < \infty)$. □

Proof of Proposition 6. Here we extend the proof of Proposition 5 to nonlinear demand settings. Assume that the seller's belief is $q_t = q \in [0, \hat{q})$ at the end of a given period t and that $p_{t+1} = \varphi(q) = p$. To apply the proof of Proposition 5 we need to verify inequality (C3). That is, we need to show that

$$\frac{1-\hat{q}}{\hat{q}} L(\hat{p}) \leq \frac{1-q}{q} L(p), \qquad \text{(C5)}$$

where $L(\cdot) = [1 - \rho_0(\cdot)]/[1 - \rho_1(\cdot)]$ is the rejection-likelihood-ratio function. Denote by $f \equiv L \circ \varphi$ the composition of the functions $L(\cdot)$ and $\varphi(\cdot)$ and let g: $[0, 1] \to \mathbb{R}$ be a function such that

$$g(q) = \frac{1-q}{q} L(\varphi(q)) = \frac{1-q}{q} f(q).$$

If we can show that $g(\cdot)$ is decreasing, then we are done. By hypothesis, we know that $f(q+h) \leq f(q)e^{4h}$ for all q and $q + h$ in $[0, 1]$. Letting $h \to 0$, we deduce that $(d/dq)\log f(q) \leq 4$ for all $q \in [0, 1]$. This in turn implies that $f'(q)/f(q) \leq 4 \leq 1/q(1-q)$ for all $q \in [0, 1]$. Noting that the derivative of $g(\cdot)$ is equal to $g'(q) = -[f(q) - q(1-q)f'(q)]/q^2$, we conclude that $g'(q) < 0$ for all $q \in [0, 1]$. □

Appendix D. Proofs of Propositions 7 and 8 and Theorems 2–4

As before, we will assume hypothesis 0. (The same analysis can be carried out for the case where hypothesis 1 holds.)

Proof of Theorem 2. By Proposition 3, the myopic price function $\varphi(\cdot)$ is nondecreasing. Because there does not exist a belief $q \in [0, 1]$ such that $\hat{p} \in \arg\max\{r_q(p),\ l \leq p \leq u\}$, we deduce that there exists an ϵ-neighborhood of $\hat{p}$ that lies outside $\varphi([0, 1])$. Therefore, the MBP coincides with CMBP(ϵ) in this case, and we are done by Theorem 3 below. □

Proof of Proposition 7. Fix $\alpha = 0$, $\delta > 0$, and apply Lemma A.1. □

Proof of Theorem 3. First, we observe that for sufficiently small ϵ, CMBP(ϵ) is a discriminative policy by construction. Now, denoting by p_t the price generated under CMBP(ϵ) at period t, we have

$$\begin{aligned}\Delta_0^{\pi}(T) &= \frac{1}{r_0(p_0^*)}\left[T r_0(p_0^*) - \sum_{t=1}^{T} \mathbb{E}_0^{\pi}[r_0(p_t)]\right] \\ &= \frac{1}{r_0(p_0^*)} \sum_{t=1}^{T} \mathbb{E}_0^{\pi}[r_0(p_0^*) - r_0(p_t)]. \qquad \text{(D1)}\end{aligned}$$

Note that when the value of q_{t-1} is near zero, we can express p_t in terms of p_0^*. To carry out this task, we first apply the implicit function theorem (IFT) to verify $\hat{\varphi}(\cdot)$ is differentiable around zero. Because $r_0(\cdot)$ and $r_1(\cdot)$ are differentiable, any unconstrained local maximizer $\nu(q)$ of $r_q(\cdot)$ has to satisfy the first-order necessary conditions for the unconstrained single-period expected profit maximization problem:

$$\eta(q, p) := q r_1'(p) + (1-q) r_0'(p) = 0,$$

where $p = \nu(q)$. Now, because $\varepsilon_0(\cdot)$ is strictly increasing, $r_0(\cdot)$ is strictly quasi-concave. Moreover, because p_0^* and p_1^* are interior points of the interval $[l, u]$, when q is sufficiently close to zero, we have $\partial\eta(q, p)/\partial p = r_q''(p) < 0$ by the strict quasi-concavity of $r_0(\cdot)$ and the second-order

conditions. Consequently, we deduce by IFT that each unconstrained local maximizer $\nu(\cdot)$ is differentiable around zero. Because $r_0(\cdot)$ is strictly quasi-concave, it has a unique global maximum. Thus, each such local maximizer $\nu(\cdot)$ satisfies $\nu(0) = p_0^*$, implying that in a neighborhood of zero we have $\hat{\varphi}(\cdot) = \nu_0(\cdot)$ for some local maximizer $\nu_0(\cdot)$. Therefore there exists $\epsilon_0 > 0$ such that $\hat{\varphi}(\cdot)$ is differentiable in the ϵ_0-neighborhood of zero, and by a Taylor series expansion of $\hat{\varphi}(\cdot)$ around zero we have the following:

$$\begin{aligned} p_t &= \hat{\varphi}(q_{t-1}) \\ &= p_0^* + \varphi'(0) q_{t-1} + C_t^{(1)} q_{t-1}^2, \end{aligned} \tag{D2}$$

where $C_t^{(1)} := \varphi''(\tilde{q})/2$ for some $\tilde{q} \in [0, q_{t-1}]$ and because $\hat{\varphi}(0) = \varphi(0) = p_0^*$ and $\hat{\varphi}'(0) = \varphi'(0)$. Note also that by properties of $\varphi(\cdot)$, it follows that $C_t^{(1)}$ is uniformly bounded for all t. Recalling (D1), we divide the expectation into two:

$$\begin{aligned} \Delta_0^\pi(T) = \frac{1}{r_0(p_0^*)} \sum_{t=1}^{T} (&\mathbb{E}_0^\pi [r_0(p_0^*) - r_0(p_t); q_{t-1} \geq \epsilon_0] \\ &+ \mathbb{E}_0^\pi [r_0(p_0^*) - r_0(p_t); q_{t-1} < \epsilon_0]). \end{aligned}$$

The first expectation can be bounded using Markov's inequality as follows:

$$\begin{aligned} \mathbb{E}_0^\pi [r_0(p_0^*) - r_0(p_t); q_{t-1} \geq \epsilon_0] &\leq (r_0(p_0^*) - \tilde{r}_0) \mathbb{P}_0^\pi (q_{t-1} \geq \epsilon_0) \\ &\leq (r_0(p_0^*) - \tilde{r}_0) \frac{\mathbb{E}_0^\pi [q_{t-1}]}{\epsilon_0}, \end{aligned}$$

where $\tilde{r}_0 := \min_{p \in [l,u]} \{r_0(p)\}$. For the second term we have by a Taylor expansion that

$$\begin{aligned} &\mathbb{E}_0^\pi [r_0(p_0^*) - r_0(p_t); q_{t-1} < \epsilon_0] \\ &\quad = \mathbb{E}_0^\pi \left[-\tfrac{1}{2} r_0''(p_0^*)(p_t - p_0^*)^2 + C_t^{(2)} (p_t - p_0^*)^3; q_{t-1} < \epsilon_0 \right] \end{aligned}$$

for suitable $C_t^{(2)}$ that is uniformly bounded because p_0^* is an interior point of $\{p \in \mathbb{R}: l \leq p \leq u \text{ and } |p - \hat{p}| \geq \epsilon\}$ and satisfies $r_0'(p_0^*) = 0$ by the first-order condition. Using this together with (D2), we arrive at the following:

$$\begin{aligned} &\mathbb{E}_0^\pi [r_0(p_0^*) - r_0(p_t); q_{t-1} < \epsilon_0] \\ &\quad \leq \mathbb{E}_0^\pi \left[-\tfrac{1}{2} r_0''(p_0^*)(\varphi'(0))^2 q_{t-1}^2 + C q_{t-1}^3 \right], \end{aligned}$$

for some finite positive constant C. Consequently, we have

$$\begin{aligned} \Delta_0^\pi(T) &\leq \left(1 - \frac{\tilde{r}_0}{r_0(p_0^*)}\right) \sum_{t=1}^{T} \frac{\mathbb{E}_0^\pi [q_{t-1}]}{\epsilon_0} \\ &\quad - \frac{r_0''(p_0^*)(\varphi'(0))^2}{2 r_0(p_0^*)} \sum_{t=1}^{T} \mathbb{E}_0^\pi [q_{t-1}^2] + C \sum_{t=1}^{T} \mathbb{E}_0^\pi [q_{t-1}^3] \\ &\leq \left(1 - \frac{\tilde{r}_0}{r_0(p_0^*)}\right) \cdot \frac{\mu}{\epsilon_0} \cdot \frac{1 - e^{-\lambda T}}{1 - e^{-\lambda}} \\ &\quad - \frac{\mu r_0''(p_0^*)(\varphi'(0))^2}{2 r_0(p_0^*)} \cdot \frac{1 - e^{-\lambda T}}{1 - e^{-\lambda}} + C \frac{1 - e^{-\lambda T}}{1 - e^{-\lambda}} \end{aligned}$$

for positive constants μ and λ given in Proposition 7 because CMBP(ϵ) is a discriminative policy for any $\epsilon > 0$. Taking $T \to \infty$, the result follows. □

Proof of Proposition 8. Assume without loss of generality that $p_0^* = \hat{p}$. To keep notation simple, let $x_t := p_t - \hat{p}$ for all t, and $x_i^* := p_i^* - \hat{p}$ for $i = 0, 1$, where $(p_1, p_2, \ldots)$ is the sequence of prices exercised by policy π. We will show that $\Delta_i^\pi(\infty)$ cannot be bounded by a constant for some $i = 0, 1$. Assume toward a contradiction that there exists a finite positive constant c_0 such that $\Delta_i^\pi(\infty) \leq c_0$ for $i = 0, 1$. Then,

$$\Delta_0^\pi(t) \leq c_0 \quad \text{for all } t = 1, 2, \ldots. \tag{D3}$$

Moreover, by definition of $\Delta_i^\pi(\cdot)$, we have

$$\Delta_0^\pi(t) = \frac{1}{r_0(p_0^*)} \sum_{k=1}^{t} \mathbb{E}_0^\pi [r_0(p_0^*) - r_0(p_k)].$$

Letting $b_0 := -\frac{1}{2} \max\{r_0''(p): p_0^* \leq p \leq p_1^*\}$, we deduce that

$$\begin{aligned} \Delta_0^\pi(t) &\geq \frac{b_0}{r_0(p_0^*)} \sum_{k=1}^{t} \mathbb{E}_0^\pi [(p_0^* - p_k)^2] \\ &= \frac{b_0}{r_0(p_0^*)} \sum_{k=1}^{t} \mathbb{E}_0^\pi [(x_0^* - x_k)^2] = \frac{b_0}{r_0(p_0^*)} \mathbb{E}_0^\pi \left[\sum_{k=1}^{t} x_k^2 \right], \end{aligned} \tag{D4}$$

because $r_0'(p_0^*) = 0$ and $x_0^* = p_0^* - \hat{p} = 0$. Denoting by $J_t := \sum_{k=1}^{t} x_k^2$ the sum of squared price deviations, we combine inequalities (D3) and (D4) to get

$$\mathbb{E}_0^\pi J_t \leq K_0 < \infty \quad \text{for all } t = 1, 2, \ldots, \tag{D5}$$

where $K_0 := c_0 r_0(p_0^*)/b_0$. Now, we will show that inequality (D5) implies $\Delta_1^\pi(\infty) = \infty$. Recalling Equation (A3), we have for all $t = 1, 2, \ldots$ the following:

$$\begin{aligned} q_t &= \frac{q_0 \prod_{k=1}^{t} \alpha_k^{X_k}}{q_0 \prod_{k=1}^{t} \alpha_k^{X_k} + \bar{q}_0 \prod_{k=1}^{t} \beta_k^{1 - X_k}} \\ &= \frac{1}{1 + (\bar{q}_0/q_0) \prod_{k=1}^{t} \alpha_k^{-X_k} \beta_k^{1 - X_k}} \\ &= \frac{1}{1 + (\bar{q}_0/q_0) \exp(\sum_{k=1}^{t} [(1 - X_k) \log \beta_k - X_k \log \alpha_k])}, \end{aligned} \tag{D6}$$

where $\rho_{i,t} = \rho_i[\pi_t(q_{t-1})]$ for all $i = 0, 1$ and $t \in \mathbb{N}$, $\alpha_k = \rho_{1,k}/\rho_{0,k}$, and $\beta_k = (1 - \rho_{0,k})/(1 - \rho_{1,k})$ for all $k \in \mathbb{N}$. Fix $q_0 \in (0, 1)$ and let L_t denote the argument of the $\exp(\cdot)$ in the denominator in (D6). Using the definition of α_k and β_k above, we get the following under the measure $\mathbb{P}_0^\pi$:

$$\begin{aligned} L_t &= \sum_{k=1}^{t} [(1 - X_k) \log \beta_k - X_k \log \alpha_k] \\ &= \sum_{k=1}^{t} \left[(\bar{\rho}_{0,k} - y_k) \log \left(\frac{\bar{\rho}_{0,k}}{\bar{\rho}_{1,k}} \right) + (\rho_{0,k} + y_k) \log \left(\frac{\rho_{0,k}}{\rho_{1,k}} \right) \right], \end{aligned}$$

where $y_k := X_k - \mathbb{E}_0^\pi X_k = X_k - \rho_{0,k}$. Because $\log z \leq z - 1$ for all $z \in \mathbb{R}_+$, the above equation implies that

$$\begin{aligned} L_t &\leq \sum_{k=1}^{t} \left[(\bar{\rho}_{0,k} - y_k) \frac{\bar{\rho}_{0,k} - \bar{\rho}_{1,k}}{\bar{\rho}_{1,k}} + (\rho_{0,k} + y_k) \frac{\rho_{0,k} - \rho_{1,k}}{\rho_{1,k}} \right] \\ &= \sum_{k=1}^{t} (\rho_{0,k} - \rho_{1,k}) \left[\frac{\rho_{0,k} \bar{\rho}_{1,k} - \rho_{1,k} \bar{\rho}_{0,k}}{\rho_{1,k} \bar{\rho}_{1,k}} - \frac{\bar{\rho}_{1,k} + \rho_{1,k}}{\rho_{1,k} \bar{\rho}_{1,k}} y_k \right] \end{aligned}$$

$$= \sum_{k=1}^{t} (\rho_{0,k} - \rho_{1,k}) \left[\frac{\rho_{0,k} - \rho_{1,k} + y_k}{\rho_{1,k}\bar{\rho}_{1,k}} \right]$$

$$\leq \sum_{k=1}^{t} \frac{\Lambda^2 x_k^2 - \lambda x_k y_k}{\rho_{1,k}\bar{\rho}_{1,k}},$$

where $\lambda := \min\{\rho_1(p) - \rho_0(p)\colon\ p_0^* \leq p \leq p_1^*\}$ and $\Lambda := \max\{\rho_1(p) - \rho_0(p)\colon\ p_0^* \leq p \leq p_1^*\}$. Moreover, putting $\gamma_0 := [\rho_{\min}(1 - \rho_{\max})]^{-1}$ and $M_t := \sum_{k=1}^{t} x_k y_k$ for all $t = 1, 2, \ldots,$ we obtain the following inequality:

$$L_t \leq \gamma_0(\Lambda^2 J_t - \lambda M_t).$$

Recalling Equation (D6), we deduce the following under the measure $\mathbb{P}_0^\pi$:

$$q_t \geq \frac{1}{1 + (\bar{q}_0/q_0)\exp(\gamma_0(\Lambda^2 J_t - \lambda M_t))}. \tag{D7}$$

With the above definition of M_t, we have $\mathbb{E}_0^\pi[M_t] = 0$ because $\mathbb{E}_0^\pi[y_t \mid x_t] = 0$ for all t. Moreover, letting $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$, we can use the same reasoning to get

$$\mathbb{E}_0^\pi[M_t - M_{t-1} \mid \mathcal{F}_{t-1}] = \mathbb{E}[x_t y_t \mid \mathcal{F}_{t-1}] = 0.$$

Putting $M_0 = 0$, we conclude that $(M_t, \mathcal{F}_t)$ is a martingale with zero mean. Let $\langle M \rangle_t$ denote the predictable compensator of $\{M_t\}$; that is,

$$\langle M \rangle_t = \sum_{k=1}^{t} \mathbb{E}_0^\pi[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}] \tag{D8}$$

for all t. Then, letting $\gamma_1 := \rho_{\max}(1 - \rho_{\min})$, we obtain

$$\begin{aligned} \langle M \rangle_t &= \sum_{k=1}^{t} x_k^2 \mathbb{E}_0^\pi[y_k^2 \mid \mathcal{F}_{k-1}] \\ &= \sum_{k=1}^{t} x_k^2 \rho_{0,k}\bar{\rho}_{0,k} \\ &\leq \gamma_1 J_t. \end{aligned} \tag{D9}$$

The lower bound on the belief process q_t, which is displayed on the right-hand side of inequality (D7), will help us show that q_t does not converge in probability to zero under the measure $\mathbb{P}_0^\pi$. Note that inequality (D5) implies the following by Markov's inequality:

$$\mathbb{P}_0^\pi(J_t \geq 3K_0) \leq \frac{\mathbb{E}_0^\pi J_t}{3K_0} \leq \frac{1}{3} \quad \text{for all } t = 1, 2, \ldots. \tag{D10}$$

Using Markov's inequality for the second time, we get

$$\begin{aligned} \mathbb{P}_0^\pi(|M_t| \geq \sqrt{3\gamma_1 K_0}) &\leq \frac{\mathbb{E}_0^\pi[M_t^2]}{3\gamma_1 K_0} \\ &\leq \frac{\gamma_1 \mathbb{E}_0^\pi[J_\infty]}{3\gamma_1 K_0} \leq \frac{1}{3}, \end{aligned} \tag{D11}$$

where the last inequalities follow from (D9) and (D5). Combining the inequalities (D10) and (D11), we deduce that

$$\mathbb{P}_0^\pi(J_t < 3K_0,\ |M_t| < \sqrt{3\gamma_1 K_0}) \geq \tfrac{1}{3} \quad \text{for all } t = 1, 2, \ldots.$$

By (D7), this implies

$$\mathbb{P}_0^\pi(q_t \geq \underline{q}) \geq \tfrac{1}{3} \quad \text{for all } t = 1, 2, \ldots, \tag{D12}$$

where

$$\underline{q} := \frac{1}{1 + (\bar{q}_0/q_0)\exp(3\Lambda^2\gamma_0 K_0 - \lambda\gamma_0\sqrt{3\gamma_1 K_0})} \in (0, 1).$$

Inequality (D12) implies that q_t does not converge in probability to zero (under the measure $\mathbb{P}_0^\pi$). As argued in the proof of Theorem 1, we deduce that π cannot charge a δ-discriminative price infinitely often because q_t does not converge to zero almost surely. Moreover, by inequality (D12), we know that π charges prices $p_t \in \{\pi_t(q)\colon\ \underline{q} \leq q \leq 1\}$ infinitely often. Thus, for any given δ, π will charge the δ-discriminative prices in the set $\{\pi_t(q)\colon\ \underline{q} \leq q \leq 1\}$ only finitely often and the non-δ-discriminative prices in the same set infinitely often. Letting δ tend to zero, we deduce that π will eventually charge the price $\pi_t(q_{t-1}) = p_0^* = \hat{p}$ for some $q_{t-1} = \tilde{q} \in [\underline{q}, 1]$ and some finite t. But then at the same time period t, q_{t-1} can attain the value $\tilde{q}$ with positive probability under the measure $\mathbb{P}_1^\pi$. Consequently, we have $\mathbb{P}_1^\pi(p_k = p_0^* \text{ for all } k \geq t) > 0$, and $\Delta_1^\pi(\infty) = \infty$. □

Proof of Theorem 4. Using the arguments in the proof of Theorem 3, we know that there exists $\epsilon_0 > 0$ such that the FMBP price function $\hat{\varphi}_t(\cdot)$ is differentiable in the ϵ_0-neighborhood of zero for all t. Therefore, by a Taylor series expansion of $\hat{\varphi}_t(\cdot)$ around zero, we get the following:

$$\begin{aligned} \Delta_0^\pi(T) \leq{}& \left(1 - \frac{\tilde{r}_0}{r_0(p_0^*)}\right) \sum_{t=1}^{T} \frac{\mathbb{E}_0^\pi[q_{t-1}]}{\epsilon_0} \\ &- \frac{r_0''(p_0^*)(\varphi'(0))^2}{2r_0(p_0^*)} \sum_{t=1}^{T} \mathbb{E}_0^\pi[q_{t-1}^2] + C \sum_{t=1}^{T} \mathbb{E}_0^\pi[q_{t-1}^3], \end{aligned}$$

where $\tilde{r}_0 := \min_{p \in [l,u]}\{r_0(p)\}$ and C is some finite positive constant. To apply Lemma A.1, we need to find some $\delta > 0$ such that FMBP($\boldsymbol{\epsilon}$) is $\boldsymbol{\delta}$-discriminative for a sequence $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots)$ that satisfies $\delta_t = \delta t^{-\alpha}$. Recalling that the function $p \mapsto \rho_1(p) - \rho_0(p)$ is locally increasing around $\hat{p}$ and vanishes at $\hat{p}$, we deduce that there exists $\xi_0 > 0$ such that the function $p \mapsto \rho_1(p) - \rho_0(p)$ is decreasing in the ξ_0-neighborhood of $\hat{p}$. Letting $B_1 := \max\{\rho_1'(p) - \rho_0'(p)\colon\ \hat{p} - \xi_0 \leq p \leq \hat{p} + \xi_0\}$, we further realize that $|\rho_1(p) - \rho_0(p)| > B_1|p - \hat{p}|$ for all $\hat{p} - \xi_0 \leq p \leq \hat{p} + \xi_0$. Because the function $p \mapsto \rho_1(p) - \rho_0(p)$ is continuous and does not vanish outside the ξ_0-neighborhood of $\hat{p}$, there exists $B_2 > 0$ such that $|\rho_1(p) - \rho_0(p)| > B_2$ for all $p \in [l, \hat{p} - \xi_0) \cup (\hat{p} + \xi_0, u]$. Fixing $\delta := \min\{B_1\xi_0, B_2\}$, we construct the sequence $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots)$ by setting $\delta_t = \delta t^{-\alpha}$ for all t. Consequently, we conclude that FMBP($\boldsymbol{\epsilon}$) is a $\boldsymbol{\delta}$-discriminative policy and get

$$\begin{aligned} \Delta_0^\pi(T) \leq{}& \left(1 - \frac{\tilde{r}_0}{r_0(p_0^*)}\right) \cdot \frac{\mu}{\epsilon_0} \sum_{t=1}^{T} \exp(-\lambda t^{1-4\alpha}) \\ &- \frac{\mu r_0''(p_0^*)(\varphi'(0))^2}{2r_0(p_0^*)} \sum_{t=1}^{T} \exp(-\lambda t^{1-4\alpha}) \\ &+ C \sum_{t=1}^{T} \exp(-\lambda t^{1-4\alpha}) \\ ={}& \left[\left(1 - \frac{\tilde{r}_0}{r_0(p_0^*)}\right) \cdot \frac{\mu}{\epsilon_0} - \frac{\mu r_0''(p_0^*)(\varphi'(0))^2}{2r_0(p_0^*)} + C \right] \\ &\cdot \sum_{t=1}^{T} \exp(-\lambda t^{1-4\alpha}) \end{aligned}$$

for positive constants μ and λ given in Lemma A.1 because $q_{t-1}^k \le q_{t-1}$ for all k. Note that

$$\sum_{t=1}^{T} \exp(-\lambda t^{1-4\alpha}) \le \int_{t=0}^{T} \exp(-\lambda t^{1-4\alpha}) dt = \int_{t=0}^{T} \exp(-\lambda t^{1/R}) dt,$$

where $R := (1-4\alpha)^{-1} \in [1, \infty)$. By a change of variable in the last integral, we get

$$\begin{aligned}\sum_{t=1}^{T} \exp(-\lambda t^{1-4\alpha}) &\le \int_{t=0}^{\lambda T^{1/R}} \exp(-u) R\lambda^{-R} u^{R-1} du \\ &= R\lambda^{-R}\gamma(R, \lambda T^{1/R}) \\ &\le R\lambda^{-R}\Gamma(R),\end{aligned}$$

where $\gamma(\cdot,\cdot)$ and $\Gamma(\cdot)$ are the lower incomplete gamma function and the gamma function, respectively. This concludes the proof. □

Appendix E. Proofs of Side Lemmas

Proof of Lemma A.2. Define h: $[0,1]^2 \to \mathbb{R}$ as

$$h(x,y) := x \log \frac{x}{y} - (1-x) \log \frac{1-x}{1-y} - 2(x-y)^2.$$

Let $x, y \in [0,1]$ and assume without loss of generality that $y \le x$. Then,

$$\frac{\partial h(x,y)}{\partial y} = \frac{y-x}{y(1-y)} - 4(y-x) \le 0$$

because $y(1-y) \le \frac{1}{4}$ and $y \le x$. Noting that $h(x,y) = 0$ when $y = x$, we deduce that $h(x,y) \ge 0$ for all $0 \le y \le x \le 1$. By symmetry, we conclude that $h(x,y) \ge 0$ for all $x, y \in [0,1]$. Finally, we note that

$$\begin{aligned}\bar{\rho}_{0,k} \log \frac{\bar{\rho}_{0,k}}{\bar{\rho}_{1,k}} + \rho_{0,k} \log \frac{\rho_{0,k}}{\rho_{1,k}} &= h(\rho_{1,k}, \rho_{0,k}) + 2(\rho_{1,k} - \rho_{0,k})^2 \\ &\ge 2(\rho_{1,k} - \rho_{0,k})^2 \ge 2\delta^2,\end{aligned}$$

where the first inequality follows because $h \ge 0$, and the second one follows because π is δ-discriminative. This concludes the proof. □

Proof of Lemma A.3. Let $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$ and recall that q_t is $\mathcal{F}_t$-measurable. Note that $\rho_{i,t} = \rho_i[\pi_t(q_{t-1})]$ is also $\mathcal{F}_t$-measurable for all $i = 0, 1$ and all t. Therefore, M_t is $\mathcal{F}_t$-measurable. Observe that

$$\begin{aligned}\mathbb{E}_0^\pi[M_t \mid \mathcal{F}_{t-1}] &= \mathbb{E}_0^\pi\left[\sum_{k=1}^{t} (X_k - \rho_{0,k}) \log \frac{\rho_{0,k}\bar{\rho}_{1,k}}{\bar{\rho}_{0,k}\rho_{1,k}} \,\middle|\, \mathcal{F}_{t-1}\right] \\ &= M_{t-1} + \mathbb{E}_0^\pi\left[(X_t - \rho_{0,t}) \log \frac{\rho_{0,t}\bar{\rho}_{1,t}}{\bar{\rho}_{0,t}\rho_{1,t}} \,\middle|\, \mathcal{F}_{t-1}\right] \\ &= M_{t-1}\end{aligned}$$

because $\mathbb{E}_0^\pi[X_t \mid \mathcal{F}_{t-1}] = \rho_{0,t}$ and $\rho_{i,t}$ is $\mathcal{F}_{t-1}$-measurable. Hence $\{M_t\}$ is an $\mathcal{F}_t$-martingale, and it is a matter of simple algebra to see that the martingale differences are bounded, in particular,

$$\begin{aligned}|M_t - M_{t-1}| &= \left|(X_t - \rho_{0,t}) \log \frac{\rho_{0,t}\bar{\rho}_{1,t}}{\bar{\rho}_{0,t}\rho_{1,t}}\right| \\ &\le \max_{p\in[l,u]} \log \rho_0(p) + \max_{p\in[l,u]} \log \bar{\rho}_1(p) \\ &\quad - \min_{p\in[l,u]} \log \bar{\rho}_0(p) - \min_{p\in[l,u]} \log \rho_1(p) \\ &\le \log\left[\frac{\rho_{\max}(1-\rho_{\min})}{\rho_{\min}(1-\rho_{\max})}\right],\end{aligned}$$

with $\rho_{\max} := \max\{\rho_0(l), \rho_1(l)\}$ and $\rho_{\min} := \min\{\rho_0(u), \rho_1(u)\}$. Denoting $\gamma := \{\log[\rho_{\max}(1-\rho_{\min})] - \log[\rho_{\min}(1-\rho_{\max})]\}^2$, we have by the Azuma–Hoeffding inequality (Williams 1991, p. 237) that

$$\mathbb{P}(|M_t| \ge t\xi) \le 2\exp\left(-\frac{1}{2\gamma}\xi^2 t\right).$$

This concludes the proof. □

References

Aghion, P., P. Bolton, C. Harris, B. Jullien. 1991. Optimal learning by experimentation. *Rev. Econom. Stud.* **58**(4) 621–654.

Anderson, T. W., J. Taylor. 1976. Some experimental results on the statistical properties of least squares estimates in control problems. *Econometrica* **44**(6) 1289–1302.

Aviv, Y., A. Pazgal. 2005. A partially observed Markov decision process for dynamic pricing. *Management Sci.* **51**(9) 1400–1416.

Besbes, O., A. Zeevi. 2009. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Oper. Res.* **57**(6) 1407–1420.

Bolton, P., C. Harris. 1999. Strategic experimentation. *Econometrica* **67**(2) 349–374.

Dembo, A., O. Zeitouni. 1998. *Large Deviations Techniques and Applications*. Springer-Verlag, New York.

Easley, D., N. M. Kiefer. 1988. Controlling a stochastic process with unknown parameters. *Econometrica* **56**(5) 1045–1064.

Farias, V. F., B. van Roy. 2010. Dynamic pricing with a prior on market response. *Oper. Res.* **58**(1) 16–29.

Girard, G. 2000. Revenue management: The price can't be right if the tools aren't. *AMR Res. Rep.* (September) 3–20.

Gittins, J. C. 1989. *Bandit Processes and Dynamic Allocation Indices*. Wiley, New York.

Keller, G., S. Rady. 1999. Optimal experimentation in a changing environment. *Rev. Econom. Stud.* **66**(3) 475–507.

Lai, T. L., H. Robbins. 1982. Iterated least squares in multiperiod control. *Adv. Appl. Math.* **3**(1) 50–73.

Levina, T., Y. Levin, J. McGill, M. Nediak. 2009. Dynamic pricing with online learning and strategic consumers: An application of the aggregating algorithm. *Oper. Res.* **57**(2) 327–341.

Lobo, M. S., S. Boyd. 2003. Pricing and learning with uncertain demand. Working paper, Stanford University, Stanford, CA.

McLennan, A. 1984. Price dispersion and incomplete learning in the long run. *J. Econom. Dynam. Control.* **7**(3) 331–347.

Morel, P., G. Stalk, P. Stanger, P. Wetenhall. 2003. Pricing myopia. Report, The Boston Consulting Group Perspectives, Boston.

Phillips, R. 2005. *Pricing and Revenue Optimization*. Stanford University Press, Stanford, CA.

Phillips, R. 2010. Private communication with J. Michael Harrison, July 2010, Stanford University, Stanford, CA.

Preslan, L., E. Newmark, R. Bois, A. Dyson. 2005. How to select the right B2B price management vendor. *AMR Res. Rep.* (May) 3–54.

Robbins, H. 1951. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58** 527–535.

Rothschild, M. 1974. A two-armed bandit theory of market pricing. *J. Econom. Theory* **9**(2) 185–202.

Talluri, K., G. van Ryzin. 2004. *The Theory and Practice of Revenue Management*. Springer, New York.

Topkis, D. M. 1978. Minimizing a submodular function on a lattice. *Oper. Res.* **26**(2) 305–321.

Williams, D. 1991. *Probability with Martingales*. Cambridge University Press, Cambridge, UK.