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

J. Michael Harrison† Stanford University

N. Bora Keskin ‡ Stanford University

Assaf Zeevi§ Columbia University

This version: January 14, 2010

Abstract

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 what we call an indeterminate equilibrium, where learning ceases prematurely and profit performance is poor. However, two variants of the myopic policy are 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. These modifications of the MBP perform so well in simulation experiments that the pursuit of an exactly optimal policy appears pointless for all practical purposes.

Keywords: Revenue management, pricing, estimation, Bayesian learning, exploration-exploitation.

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

*Research partially supported by NSF grant DMI-0447562

†Graduate School of Business, e-mail: harrison_michael@gsb.stanford.edu

‡Graduate School of Business, e-mail: keskin_bora@gsb.stanford.edu

§Graduate School of Business, e-mail: assaf@gsb.columbia.edu

seller confronts a trade-off between exploration of the demand environment (*learning*) and expected immediate profit (*earning*).

Overview of the problem and approach. The particular variant of the learning-and-earning problem that we consider here has three salient features: (a) 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; (b) the seller offers prices sequentially to individual customers, observing either success or failure in each sales attempt; and (c) the time horizon is finite but possibly large, as opposed to the infinite-horizon discounted formulations that are common in the literature of both economics and management science. Model feature (b) corresponds to what Phillips (2005) calls *customized pricing*; Chapter 11 of that book 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. It was an interest in such financial service applications that originally motivated our study.

Probably the most influential and frequently cited historical studies of learning-and-earning are those by Rothschild (1974), Easley and Kiefer (1988) and Aghion, Bolton, Harris and Jullien (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. Easley and Kiefer (1988) and Aghion et al. (1991) expanded on that finding by considering fairly general action spaces for the seller. As usual in economic theory, the authors were not primarily interested in computing optimal policies, or even in developing a modeling framework that could plausibly be used in practice, but rather in characterizing what might be called the social outcome of a particular market situation.

In contrast, we approach learning-and-earning as a management science problem, the eventual goal being to identify model structures and computational methods that are suitable for literal application. Following common practice in the revenue management industry, we think in terms of a parametric model class (such as the logit family of price-response functions, or the probit family, or the linear family), using a prior distribution over model parameters to express uncertainty about the demand environment. Also, recognizing that it will probably be impossible to determine exactly optimal pricing policies, we are interested in sub-optimal polices that have provably good properties and perform well in simulation studies.

Despite the practical aspirations declared immediately above, we restrict attention in this paper to the artificial case where one of two demand models is known to apply. To put that another way, we restrict attention to the case where the seller's prior distribution is concentrated on just two possible parameter sets, referred to hereafter as the case of a *binary prior*. The motivation

for doing this 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. Readers will see that the binary-prior version of the learn-and-earn problem is still surprisingly subtle, defying exact solution; its analysis cannot be accomplished as a quick preliminary to something more "serious." Future work will consider the case of a general (dispersed) prior distribution for initially unknown demand parameters.

Related literature. 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, cf. Chapter 5 of Talluri and van Ryzin (2004). However, unlike the vast majority of OR/MS researchers, we do not treat demand model parameters as known data.

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 (2009) is a very 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 (2009) 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 (2009) and of 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.

As indicated earlier, our work also intersects with the economics literature on price experimentation that was initiated by Rothschild (1974). That pioneering study used the multi-armed 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, cf. 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, cf. Bolton and Harris (1999), and experimentation when the demand model may change over time, cf. Keller and Rady (1999).

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. Like us, McLennan 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. Following in the footsteps of Rothschild (1974), McLennan seeks to show that the phenomenon of "incomplete learning" can occur in a dynamic pricing environment where the seller can choose among a *continuum* of prices, as opposed to the finite set of potential prices assumed in Rothschild's analysis. Like Rothschild, he adopts an infinite-horizon formulation with discounting (the discounting is crucial), and focuses on long-run behavior under the optimal Bayesian policy (that is, the policy which maximizes the expected present value of infinite-horizon profits). At the end of this section, after the content of our paper has been outlined, more will be said about relationship between our analysis and McLennan's.

Focus on myopic Bayesian policies. Following recent advances in computational methods, Bayesian inference has become the dominant approach to statistical estimation in companies that do analytical pricing. A common practice is to first estimate unknown model parameters using Bayesian methods, and then choose the optimal price given those parameter values, cf. Chapter 11 of Phillips (2005). 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 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 accumulates. 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.

Remainder of the paper. Section 2 describes in mathematical terms the problem to be addressed, including a definition of the Bayesian pricing policies to which we restrict attention. In Section 3 we show that, under any Bayesian policy, the seller's posterior beliefs (that is, the sequence of posterior probabilities assigned to the two possible demand models after successive

sales outcomes) converge to some limit almost surely, and they converge exponentially fast to a degenerate limit (that is, a limit that assigns probability 1 to a particular demand model) under what we call a *discriminative* policy. Section 4 defines and analyzes the myopic Bayesian policy (MBP), showing that it can and often does produce what we call an *indeterminate equilibrium*, where learning ceases prematurely and profit performance is poor. Finally, in Section 5 we define and analyze two intuitively appealing variants of the MBP that perform well both in theory and in simulation experiments. The proofs of all formal results are postponed to a sequence of appendices that conclude the paper.

Comparison with the work of McLennan (1984). A key feature of the binary-prior model on which we focus is the existence of a single "uninformative price" $\widehat{p}$ (see Figure 1 in Section 3). We show that, if the finite planning horizon is extended indefinitely in our model, prices may converge to $\widehat{p}$ under the MBP; this is the "indeterminate equilibrium" referred to immediately above. In a similar vein, McLennan (1984) shows that prices may converge to $\widehat{p}$ under the optimal Bayesian policy in his discounted, infinite-horizon formulation; he uses the phrase "incomplete learning" to characterize this outcome. McLennan does not determine the optimal Bayesian policy for general parameter values (that remains an unsolved problem), but still he is able to show that the parameters of his two linear demand functions can be chosen so that incomplete learning occurs with positive probability. His goal is simply to show that such negative examples exist, thus supplementing the analysis of Rothschild (1974), and he does not suggest that the sequential pricing model with probabilistic response is of any practical significance. We provide a more comprehensive theoretical framework for the Bayesian sequential pricing problem, and our focus is on the MBP and its variants, rather than on discount-optimal policies. Given the latter distinction, our analysis of "incomplete learning" is more extensive than McLennan's, and most importantly, our ultimate emphasis is on variants of the MBP that have provably good performance and are practically implementable.

2 Problem Formulation

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 $[\ell, u]$, where $0 \leq l < u < \infty$, 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 demand models $\rho_0(\cdot)$ and $\rho_1(\cdot)$ are assumed to be continuously differentiable and strictly decreasing over $[\ell, u]$, 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 $[\ell, u]$. We further assume that p_0^* and p_1^* are interior points of the feasible price range $[\ell, u]$, and without loss of generality that $p_0^* < p_1^*$. The first-order conditions for optimality then give the following:

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

Pricing policies 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:

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

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

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.

Induced probabilities and performance metrics. 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 sales realization $x_1, x_2, \ldots$ 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 $\mathbb{P}_i^\pi(\cdot)$.

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

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

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:

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

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 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 non-decreasing as a function of T, the limit $\Delta_i^\pi(\infty)$ necessarily exists, although it may be infinite.

3 Belief Convergence

Let π be a fixed but arbitrary pricing policy. From the definition (5) it is easy to verify that the corresponding posterior probabilities $\{q_t\}$ form a bounded non-negative supermartingale under $\mathbb{P}_0^\pi$, and form a bounded, non-negative 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 a standard result in martingale theory gives the following conclusion, cf. Williams (1991, p. 109).

Proposition 1 (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 $\mathbb{P}_1^\pi$.*

Our main focus in this paper is on pricing policies π that "reveal the truth" almost surely, meaning that $\mathbb{P}_0^\pi\{q_t \to 0\} = \mathbb{P}_1^\pi\{q_t \to 1\} = 1$. However, the following analysis shows that an arbitrary policy need not have that property. In this development a price $\widehat{p} \in [\ell, u]$ is said to be *uninformative* if $\rho_0(\widehat{p}) = \rho_1(\widehat{p})$. 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 $\rho(p) = a - bp$, and in the second one they both have the logit form $\rho(p) = [1 + \exp(a + bp)]^{-1}$.

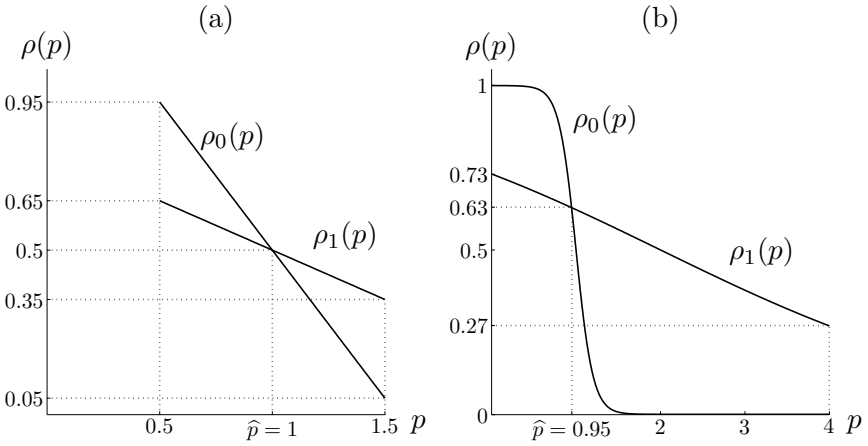

Figure 1: **Two examples of demand models from standard parametric families.** Panel (a) depicts two linear demand models, each of which represents the take-up probability of a typical customer at different price levels. Panel (b) shows two logit demand models to illustrate the customer behavior in a different setting.

To be specific, the data for the linear example are

$$\ell = 0.5, \quad u = 1.5, \quad \rho_0(p) = 1.4 - 0.9p, \quad \text{and} \quad \rho_1(p) = 0.8 - 0.3p, \tag{9}$$

and for the logit example they are

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

In each of these examples there is a unique uninformative price $\widehat{p}$ (that is, a unique price $\widehat{p}$ at which the two curves cross). Suppose that $0 < q_0 < 1$ and that the seller chooses price $p_t = \widehat{p}$ in every period. Then $\rho_0(p_t) = \rho_1(p_t)$ in every period t, so formula (5) gives $q_t = q_{t-1}$ in every period t, and obviously $q_\infty = q_0$. This is an elementary example of a belief limit in which the seller remains forever uncertain about which demand hypothesis is true. In the next section more interesting examples will be discussed.

We now show that, by avoiding uninformative prices, certain policies do more than "reveal the truth" almost surely: they find it exponentially fast. In this development a policy π is said to be *δ-discriminative* if

$$|\rho_0(\pi_t(q)) - \rho_1(\pi_t(q))| > \delta \text{ for all } t = 1, 2, \ldots \text{ and } q \in [0, 1], \tag{11}$$

and to be *discriminative* if it is δ-discriminative for some $\delta > 0$. The following proposition provides the key ingredient to establishing both of our main results (see Section 5).

Proposition 2 (rate of learning) *If π is a discriminative policy, then there exist constants μ, $\lambda > 0$ such that*

$$\mathbb{E}_0^\pi(q_t) \le \mu e^{-\lambda t} \quad \textit{and} \quad \mathbb{E}_1^\pi(1 - q_t) \le \mu e^{-\lambda t} \textit{ for all } t = 1, 2, \ldots \tag{12}$$

To conclude this section, we show 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. In stark contrast, it will be shown later that for certain discriminative policies, the expected profit loss can be bounded by a constant not depending on the horizon length T.

Proposition 3 (profit loss due to incomplete learning) *Suppose that, for a given policy π and a given hypothesis $i \in \{0, 1\}$, the limit belief q_∞ is neither almost surely 0 nor almost surely 1, and that $p_i^* \neq \widehat{p}$. Then there exists a constant $\theta > 0$ such that $\Delta_i^\pi(T) \ge \theta T$ for all T.*

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

$$r_q(p) := q r_1(p) + (1 - q) r_0(p). \tag{13}$$

For some values of q the function $r_q(\cdot)$ may achieve its maximum at multiple prices; readers will see shortly that this occurs in the logit example (10), specifically for $q = 0.54$. To resolve that ambiguity we define the *myopic price*

$$\varphi(q) := \sup \operatorname{argmax}\{r_q(p), \ell \le p \le u\} \text{ for } 0 \le q \le 1. \tag{14}$$

Recall that in Section 1 we defined $p_0^* := \varphi(0)$ and $p_1^* := \varphi(1)$, observing that the maximizing price in (14) *is* unique if $q = 0$ or $q = 1$.

Proposition 4 *$\varphi(\cdot)$ is non-decreasing on $[0, 1]$.*

The *myopic Bayesian policy* (MBP) is the pricing policy π having $\pi_t(q) = \varphi(q)$ 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 Section 1. One might plausibly hope that learning will somehow "take care of itself" under the MBP, but the analysis below shows that hope to be unfounded.

Proposition 5 (existence of indeterminate equilibrium) *There exists at most one uninformative price $\widehat{p} \in (p_0^*, p_1^*)$, and if such a $\widehat{p}$ exists, there is at most one* confounding belief *$\widehat{q} \in [0, 1]$ such that $\varphi(\widehat{q}) = \widehat{p}$.*

Hereafter a pair $(\widehat{p}, \widehat{q})$ meeting the specifications in Proposition 5 will be called an *indeterminate equilibrium* for the MBP, with the following justification: if the seller enters period t with belief $q_t = \widehat{q}$, then the MBP dictates price $p_t = \varphi(\widehat{q}) = \widehat{p}$, and because $\rho_0(\widehat{p}) = \rho_1(\widehat{p})$, formula (5) simply gives $q_{t+1} = q_t = \widehat{q}$, and the same process repeats in every subsequent period; belief convergence is to $q_\infty = \widehat{q}$. Thus the seller never learns which demand hypothesis is true, and expected profit under the MBP is strictly sub-optimal in every period.

For our linear example (9) the function $\varphi(\cdot)$ is continuous, as shown in Figure 2, and the confounding belief that corresponds to the uninformative price $\widehat{p} = 1$ (see Figure 1, left panel) is $\widehat{q} = 0.67$.

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

Proposition 6 (nature of incomplete learning) *Suppose that $\rho_0(\cdot)$ and $\rho_1(\cdot)$ are both linear, and that there exists an indeterminate equilibrium $(\widehat{p}, \widehat{q})$ for the MBP. If $q_0 \le \widehat{q}$ (respectively, $q_0 \ge \widehat{q}$), then $q_t \le \widehat{q}$ (respectively, $q_t \ge \widehat{q}$) for all $t = 1, 2, \ldots$, where $\{q_t\}$ is the seller's belief process under the MBP.*

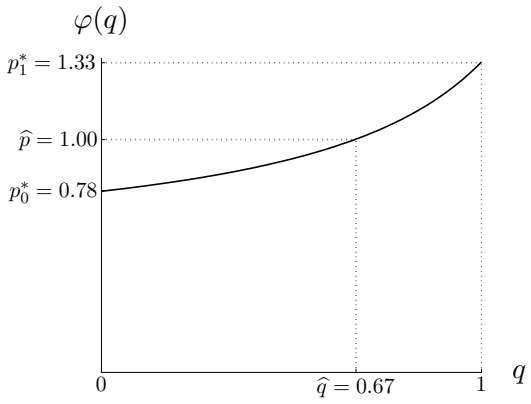

Figure 2: **Myopic price function for the linear example.** The myopic price function $\varphi(\cdot) : [0,1] \to [\ell, u]$ in the linear example increases *continuously* over its domain.

A simple verbal paraphrase of Proposition 6 is the following: Under the MBP, a seller's belief process $\{q_t\}$ cannot jump over the confounding belief $\widehat{q}$.

Proposition 6 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 $q_0 \leq \widehat{q}$. Our starting point is the following assertion: in light of Propositions 5 and 6, the only possible value for q_∞ under hypothesis 1 is $\widehat{q}$, and the only possible values under hypothesis 0 are $\widehat{q}$ and 0. The proof of that assertion is left as an exercise.

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

$$\begin{aligned} q_0 = \mathbb{E}(q_\infty) &= q_0\mathbb{E}_1(q_\infty) + (1-q_0)\mathbb{E}_0(q_\infty) \\ &= q_0\widehat{q} + (1-q_0)\widehat{q}\,\mathbb{P}_0(q_\infty = \widehat{q}). \end{aligned} \tag{15}$$

Solving this equation gives

$$\mathbb{P}_0(q_\infty = \widehat{q}) = \frac{q_0(1-\widehat{q})}{\widehat{q}(1-q_0)}. \tag{16}$$

A virtually identical calculation gives the distribution of q_∞ when $q_0 \geq \widehat{q}$, and readers will see that these calculations are equally valid for any other example where: (a) an indeterminate equilibrium $(\widehat{p}, \widehat{q})$ exists for the MBP; and (b) the seller's belief process cannot jump over $\widehat{q}$. In the numerical experiments that we have undertaken with demand models from various parametric families, such examples have proved to be the rule rather than the exception, but for the logit example specified

in (10) and pictured in the right panel of Figure 1, *there does not exist a confounding belief* $\widehat{q}$. In fact, for reasons explained in the next paragraph, the MBP is a discriminative policy in our logit example, so Proposition 2 ensures that it "reveals the truth" exponentially fast.

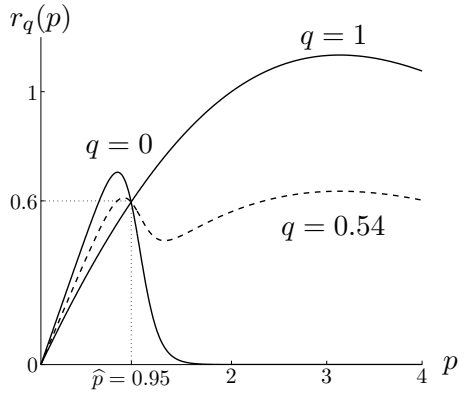

Figure 3: **Single-period expected profit function for the logit example.** The expected profit function $r_q(\cdot)$ is unimodal when q takes the extreme values of 0 and 1. It has however two global maxima when $q = 0.54$.

Figure 3 shows the single-period expected profit function $r_q(\cdot)$ for three different values of q in our logit example (10). 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 higher for $q < 0.54$, the second one is higher for $q > 0.54$, and they are equally high when $q = 0.54$. As a consequence, the myopic price $\varphi(q)$ has the behavior pictured in Figure 4, jumping from 0.87 to 3.12 as q passes through the critical value of 0.54. The unique uninformative price for this logit model is $\widehat{p} = 0.95$, and it lies in the interval of values over which $\varphi(\cdot)$ jumps, so the MBP is discriminative, as stated previously.

Taking the seller's pricing policy π to be the MBP, let us 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 assumed. Figure 5 plots estimates of $\Delta(T)$ for various horizon lengths T, both for the linear example (9) and for the logit example (10). These estimates are based on 10,000 independently generated sales sequences, taking the seller's prior belief to be $q_0 = 0.5$. Note that hypothesis 1 is true in half of these simulations, and $\widehat{q} > 0.5$ for our linear example. Thus it follows from Proposition 6 that $q_t \to \widehat{q}$ almost surely as $t \to \infty$, when hypothesis 1 is assumed in the linear example. On the other hand, as explained earlier, even under hypothesis 0 one observes that $q_t \to \widehat{q}$ as $t \to \infty$ with positive probability. Hence in the linear example the MBP stops learning with positive probability and the cumulative loss $\Delta(\cdot)$ grows linearly with the horizon (as asserted in Proposition 3). In contrast, because the MBP is a discriminative policy in the logit example, it follows from Proposition 2 that beliefs converge to the true underlying hypothesis exponentially

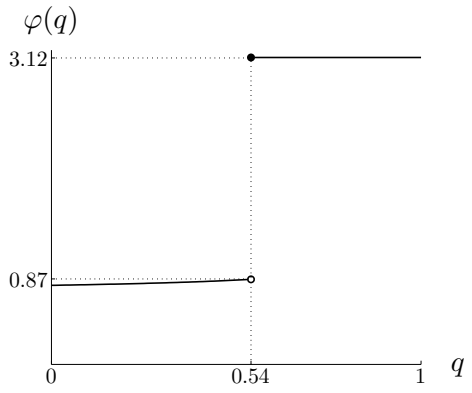

Figure 4: **Myopic price function for the logit example.** The myopic price $\varphi(\cdot)$ has a jump discontinuity in the logit example. Since the discontinuity gap contains the uninformative price $\widehat{p} = 0.95$, there exists no confounding belief $\widehat{q}$ that satisfies $\varphi(\widehat{q}) = \widehat{p}$.

fast. In that case, a simple consequence of Theorem 2 below is that $\Delta(\cdot)$ is bounded by a constant (independent of the time horizon), which is consistent with what one sees in the right panel of Figure 5.

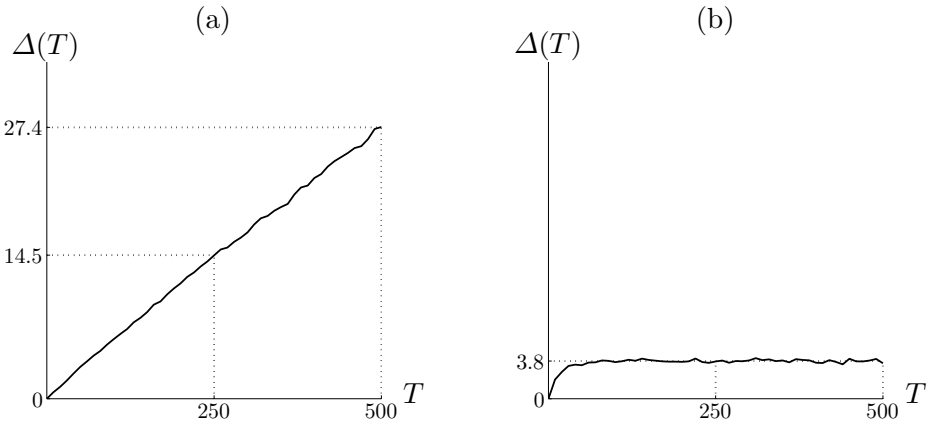

Figure 5: **Performance of the MBP in two examples.** Panel (a) exhibits the revenue loss $\Delta(T)$ of MBP, which is seen to increase linearly in the time horizon T. However, as depicted in Panel (b), the same performance metric for MBP is uniformly bounded in the logit example.

To summarize, we have shown in this section that the MBP can "get stuck" at an indeterminate equilibrium, regardless of the prior belief with which the seller begins. Such behavior is by no means exceptional, and therefore one should consider potential modifications of MBP to preclude such behavior. This is the subject of the subsequent section.

5 Two Modifications of MBP

In this section we define and analyze two modifications of the MBP that eliminate the potential for convergence to an indeterminate equilibrium. The first approach is simply to preclude prices that are in the neighborhood of the uninformative one: assuming that an indeterminate equilibrium $(\widehat{p}, \widehat{q})$ exists, and given a policy parameter $\epsilon > 0$, let

$$\widehat{\varphi}(q) := \sup \operatorname{argmax}\{r_q(p) \ : \ \ell \leq p \leq u \ \text{ and } |p - \widehat{p}| \geq \epsilon\} \tag{17}$$

for $0 \leq q \leq 1$. The pricing policy π that has $\pi_t(\cdot) = \widehat{\varphi}(\cdot)$ for all $t = 1, 2, \ldots$ is called the *constrained* variant of the MBP, with associated *threshold parameter* ϵ, hereafter abbreviated CMBP(ϵ). The following theorem says that the profit loss under CMBP(ϵ), relative to the profit earned by a clairvoyant, is bounded by a constant. The result is fairly obvious as the modification above ensures that MBP is discriminative, for which Proposition 2 guarantees exponential convergence of beliefs to the true underlying hypothesis.

Theorem 1 (performance of constrained MBP) *Assume that an indeterminate equilibrium $(\widehat{p}, \widehat{q})$ exists, and that $\epsilon > 0$ is small enough to ensure that the argument set in (17) is non-empty. 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 $\pi =$ CMBP(ϵ).*

The second modification of the MBP that we consider involves both a threshold parameter $\epsilon > 0$ and an *experimental price* $\widetilde{p} \in [\ell, u]$; we call it the *adaptive* variant of the MBP, abbreviated AMBP($\epsilon, \widetilde{p}$). Given said tuning parameters, put

$$\widetilde{\varphi}(q) = \begin{cases} \varphi(q) & \text{if } |q - \widehat{q}| \geq \epsilon \\ \widetilde{p} & \text{otherwise,} \end{cases} \tag{18}$$

for $0 \leq q \leq 1$. Then AMBP($\epsilon, \widetilde{p}$) is the pricing policy π that has $\pi_t(\cdot) = \widetilde{\varphi}(\cdot)$ for all $t = 1, 2, \ldots$ That is, AMBP($\epsilon, \widetilde{p}$) agrees with the MBP when the current belief q differs by at least ϵ from the confounding belief $\widehat{q}$, but if $|q - \widehat{q}|$ is small, then a price experiment is undertaken to ensure learning. The following analog of Theorem 1 also hinges on Proposition 2.

Theorem 2 (performance of adaptive MBP) *Assume that an indeterminate equilibrium $(\widehat{p}, \widehat{q})$ exists, that $\widetilde{p} \neq \widehat{p}$, and that $\epsilon > 0$ is small enough to ensure that $\widehat{q} \in (\epsilon, 1 - \epsilon)$. 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 $\pi =$ AMBP($\epsilon, \widetilde{p}$).*

Remark 1 The constants in both theorems are explicitly identified in the proofs, and are given in terms of the problem primitives and the tuning parameters of the 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

Tables 1 and 2 illustrate the performance of CMBP(ϵ) and AMBP($\epsilon, \widetilde{p}$) in the linear example (9). To be more specific, Tables 1 and 2 give simulation estimates for the metric $\Delta(\cdot) = \frac{1}{2}\Delta_0^\pi(\cdot) + \frac{1}{2}\Delta_1^\pi(\cdot)$ under CMBP(ϵ) and AMBP($\epsilon, \widetilde{p}$), respectively; all of these simulation estimates are based on 100,000 independent replications of the relevant sales sequence. For CMBP(ϵ) we tabulate $\Delta(T)$ for different values of the threshold 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 $\epsilon = 0.2$. In the case of AMBP($\epsilon, \widetilde{p}$), we simply fix the horizon length at T = 2,000, tabulating Δ(2,000) for different values of ϵ and $\widetilde{p}$; the optimal parameter values are seen to be $\epsilon = 0.3$ and $\widetilde{p} = 0.5$, and the corresponding performance metric Δ(2,000) = 4.1. Thus, under AMBP($\epsilon, \widetilde{p}$) with the tuning parameters optimized, the total expected loss due to initial model uncertainty equals the expected profit from just 4-5 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: Δ(2,000) under AMBP($\epsilon, \widetilde{p}$) for various values of ϵ and $\widetilde{p}$

ϵ	$\widetilde{p} = 0.50$	$\widetilde{p} = 0.67$	$\widetilde{p} = 0.83$	$\widetilde{p} = 1.00$	$\widetilde{p} = 1.17$	$\widetilde{p} = 1.33$	$\widetilde{p} = 1.50$
0.05	15.1	21.0	26.8	113.0	26.8	21.0	15.7
0.10	9.9	12.3	15.8	123.7	16.1	12.1	10.0
0.15	7.3	8.8	12.6	135.2	12.8	8.8	7.3
0.20	5.4	7.0	11.5	144.4	11.3	7.3	5.7
0.25	4.6	6.5	11.9	144.3	11.4	6.6	5.0
0.30	4.1	7.1	14.6	144.2	12.2	6.4	4.6
0.32	4.3	7.9	17.5	144.0	12.9	6.9	4.5

Theorems 1 and 2 establish that the constrained and adaptive variants of the myopic Bayesian policy eliminate its most glaring deficiency, namely, potential convergence to an indeterminate equilibrium. It is natural to ask whether these policy classes can be extended or generalized beyond the simple setting of a binary prior distribution. The CMBP family does not extend in

any obvious way, but in future work we plan to develop a generalized version of AMBP, taking as our point of departure the following observation: a policy in the AMBP family undertakes a price experiment whenever "learning slows down," and that phrase can be given precise mathematical meaning in a general setting.

A Proof of Propositions

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 Section C. 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 $\overline{y} := 1 - y$ 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 $k \in \mathbb{N}$.

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

$$q_t = \sum_{x \in \{0,1\}^t} \frac{q_0 A_t(x)}{q_0 A_t(x) + (1 - q_0) B_t(x)} \, I_{\{(X_1, \ldots, X_t) = x\}} \tag{19}$$

where $A_t(x) := \prod_{k=0}^{t-1} \alpha_k^{x_{k+1}}$ and $B_t(x) := \prod_{k=0}^{t-1} \beta_k^{1-x_{k+1}}$.

Fix $q_0 \in [0, 1)$ and fix a δ-discriminative policy π. Taking an expectation of the above expression yields:

$$\begin{aligned}
\mathbb{E}_0^\pi[q_t] &= \sum_{x \in \{0,1\}^t} \frac{q_0 A_t(x)}{q_0 A_t(x) + (1 - q_0) B_t(x)} \, \mathbb{P}_0^\pi\big[(X_1, \ldots, X_t) = x\big] \\
&= \mathbb{E}_0^\pi \left[\frac{q_0 \prod_{k=0}^{t-1} \alpha_k^{X_{k+1}}}{q_0 \prod_{k=0}^{t-1} \alpha_k^{X_{k+1}} + \overline{q}_0 \prod_{k=0}^{t-1} \beta_k^{1-X_{k+1}}} \right] \\
&= \mathbb{E}_0^\pi \left[\frac{1}{1 + \frac{\overline{q}_0}{q_0} \prod_{k=0}^{t-1} \beta_k^{1-X_{k+1}} \alpha_k^{-X_{k+1}}} \right] \\
&= \mathbb{E}_0^\pi \left[\frac{1}{1 + \frac{\overline{q}_0}{q_0} \exp\Big(\sum_{k=0}^{t-1} [(1 - X_{k+1}) \log \beta_k - X_{k+1} \log \alpha_k] \Big)} \right].
\end{aligned} \tag{20}$$

Let L_t denote the argument of the $\exp(\cdot)$ in the denominator in (20). Simple algebra, and using

the definition of α_k, β_k above, then yields

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

The following lemma allows to bound L_t from below.

Lemma A.1 *For any δ-discriminative policy π*

$$\overline{\rho}_{0,k} \log \frac{\overline{\rho}_{0,k}}{\overline{\rho}_{1,k}} + \rho_{0,k} \log \frac{\rho_{0,k}}{\rho_{1,k}} \geq 2\delta^2 \quad \textit{for all } k = 1, 2, \ldots$$

Combining the expression derived above for L_t with this inequality, we have:

$$L_t \;\geq\; \sum_{k=0}^{t-1} \left[(X_{k+1} - \rho_{0,k}) \log \frac{\rho_{0,k}\, \overline{\rho}_{1,k}}{\overline{\rho}_{0,k}\, \rho_{1,k}}\right] + 2t\delta^2.$$

We now turn to analyzing the sum on the right-hand-side above. To that end, put $M_1 = 0$ and

$$M_t := \sum_{k=1}^{t-1} \left[(X_{k+1} - \rho_{0,k}) \log \frac{\rho_{0,k}\, \overline{\rho}_{1,k}}{\overline{\rho}_{0,k}\, \rho_{1,k}}\right], \quad \text{for all } t = 2, 3, \ldots$$

Fix $\epsilon > 0$, and consider the event

$$\mathcal{J}_{t,\epsilon} := \{|M_t| < t\epsilon\}.$$

Using the above and recalling (20), we deduce that

$$\begin{aligned}
\mathbb{E}_0^\pi[q_t] &= \mathbb{E}_0^\pi\left[\frac{1}{1 + \frac{\overline{q}_0}{q_0} \exp(L_t)}\right] \\
&\leq \mathbb{E}_0^\pi\left[\frac{1}{1 + \frac{\overline{q}_0}{q_0} \exp(M_t + 2t\delta^2)}\right] \\
&= \mathbb{E}_0^\pi\left[\frac{1}{1 + \frac{\overline{q}_0}{q_0} \exp(M_t + 2t\delta^2)}; \mathcal{J}_{t,\epsilon}\right] + \mathbb{E}_0^\pi\left[\frac{1}{1 + \frac{\overline{q}_0}{q_0} \exp(M_t + 2t\delta^2)}; \mathcal{J}_{t,\epsilon}^c\right] \\
&\leq \frac{1}{1 + \frac{\overline{q}_0}{q_0} \exp(-t\epsilon + 2t\delta^2)} + \mathbb{P}_0^\pi\big(\mathcal{J}_{t,\epsilon}^c\big),
\end{aligned}$$

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

Lemma A.2 *For any policy π, there exists a positive real number γ such that*

$$\mathbb{P}_0^\pi\big(\mathcal{J}_{t,\epsilon}^c\big) \le 2\exp\big(-\frac{1}{2\gamma}\epsilon^2 t\big) \quad \textit{for all } t = 1, 2, \ldots$$

Combining this with the above, we have that

$$\begin{aligned}
\mathbb{E}_0^\pi[q_t] &\le \frac{1}{1 + \frac{\overline{q}_0}{q_0}\exp(-t\epsilon + 2t\delta^2)} + 2\exp\left[-\frac{1}{2\gamma}\epsilon^2 t\right] \\
&\le \frac{q_0}{\overline{q}_0}\exp\big[-t(2\delta^2 - \epsilon)\big] + 2\exp\left[-\frac{1}{2\gamma}\epsilon^2 t\right] \\
&\le \mu e^{-\lambda(\epsilon)t},
\end{aligned}$$

where $\mu = 2\max\{2, \frac{q_0}{\overline{q}_0}\}$ and $\lambda(\epsilon) = \min\big\{2\delta^2 - \epsilon, \frac{1}{2\gamma}\epsilon^2\big\}$. Setting $\epsilon = \delta^2 > 0$, we have $\lambda \equiv \lambda(\delta^2) = \delta^2 \min\big\{1, \delta^2/2\gamma\big\} > 0$. This completes the proof. ■

Proof of Proposition 3. 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 towards a contradiction that $\mathcal{D}_{t,\delta}^\pi$ 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.1 and direct application of Lemma A.2 leads to

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

As argued in the proof of Proposition 2 we conclude that there exist constants $\mu, \lambda > 0$ such that $\mathbb{E}_0^\pi[q_{t(k)}] \le \mu e^{-\lambda k}$. Hence, $q_\infty = 0$ almost surely, which contradicts the fact that q_∞ is neither 0 nor 1, almost surely. Therefore $\mathcal{D}_{t,\delta}^\pi$ 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 $D_\infty < \infty$ almost surely.

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

$$\begin{aligned}
\liminf_{T\to\infty} \frac{\Delta_0^\pi(T)}{T} &\ge \mathbb{E}_0^\pi\left[\liminf_{T\to\infty}\left(1 - \frac{\sum_{t=1}^T r_0(p_t)}{T r_0(p_0^*)}\right)\right] \\
&\ge \mathbb{E}_0^\pi\left[\liminf_{T\to\infty}\left(1 - \frac{D_T\, r_0(p_0^*) + (T - D_T)\widehat{r}_0}{T r_0(p_0^*)}\right)\right] \\
&= 1 - \frac{\widehat{r}_0}{r_0(p_0^*)}, \qquad (21)
\end{aligned}$$

since $D_\infty < \infty$ almost surely. The proof is complete by setting $\theta := 1 - \frac{\widehat{r}_0}{r_0(p_0^*)}$. ■

Proof of Proposition 4. Since $\varepsilon_0(\cdot)$ 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 $[\ell, p_0^*) \cap [\ell, p_1^*) = [\ell, p_0^*)$ 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, $\varphi\big([0,1]\big) \subseteq [p_0^*, p_1^*]$. 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 $r_i'(p)$. Since $\varepsilon_i(p_i^*) = 1$ 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 $\varphi(q) \in \operatorname{argmax}\{r_q(p) \,:\, \ell \le p \le u\}$ for all $q \in [0,1]$, and note that

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

Since $r_1'(p) > 0$ and $r_0'(p) < 0$ for all $p \in [p_0^*, p_1^*]$, one has $\frac{\partial^2 r_q(p)}{\partial p \partial q} > 0$ 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 Topkis's (1978) Theorem, we deduce that $\varphi(q)$ is non-decreasing in q. ■

Proof of Proposition 5. Let $\widehat{p}$ be an uninformative price. First, since $\varepsilon_1(p) < 1 < \varepsilon_0(p)$ for all $p \in [p_0^*, p_1^*]$ and $\rho_1(\widehat{p}) = \rho_0(\widehat{p})$, we deduce that $\rho_1'(\widehat{p}) - \rho_0'(\widehat{p}) > 0$. Thus, the function $p \mapsto \rho_1(p) - \rho_0(p)$ is locally increasing in the neighborhood of any uninformative price, and vanishes at that point. This implies that there can be at most one uninformative price $\widehat{p}$. Now, note that $\widehat{p} = \varphi(\widehat{q}) = \sup \operatorname{argmax}\{r_{\widehat{q}}(p) \,:\, \ell \le p \le u\}$. Thus, by the first-order conditions of optimality,

$$\widehat{q} = -\frac{\rho_0(\widehat{p}) + \widehat{p}\rho_0'(\widehat{p})}{\widehat{p}\big[\rho_1'(\widehat{p}) - \rho_0'(\widehat{p})\big]}. \tag{22}$$

This completes the proof. ■

Proof of Proposition 6. We have the existence and uniqueness of the pair $(\widehat{p}, \widehat{q})$ by Proposition 5. Moreover, as argued in the proof of Proposition 5, the function $p \mapsto \rho_1(p) - \rho_0(p)$ is locally increasing around $\widehat{p}$, and vanishes at $\widehat{p}$. Thus,

$$\rho_0(p) > \rho_1(p) \quad \forall p < \widehat{p}, \qquad \text{and} \qquad \rho_0(p) < \rho_1(p) \quad \forall p > \widehat{p}. \tag{23}$$

Without loss of generality, assume that the seller's belief is $q_t = q \in [0, \widehat{q})$ 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 $\widehat{q}$.

Case 1. $X_{t+1} = 1$. We know by Proposition 4 that the myopic price $\varphi(q)$ is monotone increasing in q. Therefore, $p = \varphi(q) \le \widehat{p}$ since $q \le \widehat{q}$. Now, by plugging $X_{t+1} = 1$ into (5), we have q_{t+1} expressed in terms of $q_t = q$ and $p_{t+1} = p$,

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

Since $p \le \widehat{p}$, by (23) we have that $(1-q)\rho_1(p) < (1-q)\rho_0(p)$. Thus, $q_{t+1}(q,p) < q \le \widehat{q}$.

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

$$q_{t+1}(q,p) = \frac{q\overline{\rho}_1(p)}{q\overline{\rho}_1(p) + (1-q)\overline{\rho}_0(p)} \leq \widehat{q},$$

or equivalently,

$$\widehat{q}(1-q)\overline{\rho}_0(p) - q(1-\widehat{q})\overline{\rho}_1(p) \geq 0. \tag{24}$$

In the case of of a linear demand model, the uninformative price and the myopic price can be explicitly expressed as $\widehat{p} = (a_0 - a_1)/(b_0 - b_1)$ 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 (22) that

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

Now, we plug the above expressions of $p = \varphi(q)$ and $\widehat{q}$ into the left-hand-side of (24), and deduce by simple algebra that (24) is equivalent to

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

To show that the above statement is correct, we recall that p_0^* and p_1^* are interior points of the interval $[\ell, u]$. We also know that $\rho_0(p_0^*) \leq 1$ and $\overline{\rho}_1(p_1^*) \leq 1$, 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 non-negative for $q < \widehat{q}$, and hence the MBP belief process q_t cannot "jump over" the confounding belief $\widehat{q}$.

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

B Proof of Theorems 1 and 2

As before, we will assume hypothesis 0 and show that $\Delta_0(\infty) < \infty$. (The same analysis can be carried out for the case where hypothesis 1 holds.)

Proof of Theorem 1. 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

$$\begin{aligned}
\Delta_0(T) &= \frac{1}{r_0(p_0^*)}\left[T\, r_0(p_0^*) - \sum_{t=1}^{T} \mathbb{E}_0^{\pi}[r_0(p_t)]\right] \\
&= \frac{1}{r_0(p_0^*)} \sum_{t=1}^{T} \mathbb{E}_0^{\pi}\big[r_0(p_0^*) - r_0(p_t)\big]. \qquad (25)
\end{aligned}$$

Note that, when the value of q_{t-1} is near 0, 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 $\widehat{\varphi}(\cdot)$ is differentiable around 0. Since

$r_0(\cdot)$ 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:

$$\eta(q,p) := qr_1'(p) + (1-q)r_0'(p) = 0,$$

where $p = \nu(q)$. Now, since $\varepsilon_0(\cdot)$ is strictly increasing, $r_0(\cdot)$ is strictly quasi-concave. Moreover, since p_0^* and p_1^* are interior points of the interval $[\ell, u]$, then when q is sufficiently close to 0, 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 0. Since $r_0(\cdot)$ 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 0 we have $\widehat{\varphi}(\cdot) = \nu_0(\cdot)$ for some local maximizer $\nu_0(\cdot)$. Therefore there exists $\epsilon_0 > 0$ such that $\widehat{\varphi}(\cdot)$ is differentiable in the ϵ_0-neighborhood of 0, and by a Taylor series expansion of $\widehat{\varphi}(\cdot)$ around 0 we have:

$$\begin{aligned} p_t &= \widehat{\varphi}(q_{t-1}) \\ &= p_0^* + \varphi'(0)\, q_{t-1} + C_t^{(1)} q_{t-1}^2, \end{aligned} \tag{26}$$

where $C_t^{(1)} := \varphi''(\widetilde{q})/2$ for some $\widetilde{q} \in [0, q_{t-1}]$, and using the fact that $\widehat{\varphi}(0) = \varphi(0) = p_0^*$ and $\widehat{\varphi}'(0) = \varphi'(0)$. Note also that by properties of $\varphi(\cdot)$, it follows that $C_t^{(1)}$ is uniformly bounded for all t. Recalling (25), we divide the expectation into two:

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

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

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

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

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

for suitable $C_t^{(2)}$ which is uniformly bounded, where we have used the fact that p_0^* is an interior point of $\{p \in \mathbb{R}\ :\ \ell \leq p \leq u \text{ and } |p - \widehat{p}| \geq \epsilon\}$ and satisfies $r_0'(p_0^*) = 0$ by the first-order condition. Using this together with (26), we arrive at:

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

for some finite positive constant C. Consequently, we have

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

for positive constants μ and λ given in Proposition 2, since CMBP(ϵ) is a discriminative policy for any $\epsilon>0$, and $q_{t-1}^k\leq q_{t-1}$ for all k. Taking $T\to\infty$ the result follows. ■

Proof of Theorem 2. The proof follows that of Theorem 1. We first observe that AMBP($\epsilon,\widetilde{p}$) is a discriminative policy for sufficiently small $\epsilon>0$ and $\widetilde{p}\neq\widehat{p}$. This follows since within an ϵ-neighborhood of $\widehat{q}$ AMBP($\epsilon,\widetilde{p}$) charges $\widetilde{p}$, which is an informative price. On the other hand, outside that neighborhood, it charges the MBP price. Then, exactly as in the proof of Theorem 1, we find an ϵ_0-neighborhood of 0 in which AMBP($\epsilon,\widetilde{p}$) charges the MBP price, and $\Delta_0(\infty)<\infty$ follows in the same manner. This concludes the proof. ■

C Proof of Side Lemmas

Proof of Lemma A.1. 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 $y\leq x$. Then,

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

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

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

where the first inequality follows since $h\geq 0$, and the second one follows since π is δ-discriminative. This concludes the proof. ■

Proof of Lemma A.2. 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

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

since $\mathbb{E}_0^\pi[X_{t+1}|\mathcal{F}_t] = \rho_{0,t}$ and $\rho_{i,t}$ is $\mathcal{F}_t$-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

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

with $\rho_{\max} := \max\{\rho_0(\ell),\rho_1(\ell)\}$, 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 Azuma's inequality (Williams 1991, p. 237) that

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

This concludes the proof. ■

Acknowledgment. We 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 our research.

References

Aghion, P., Bolton, P., Harris, C. and Jullien, B. (1991), 'Optimal Learning by Experimentation', *The Review of Economic Studies* **58**(4), 621–654.

Aviv, Y. and Pazgal, A. (2005), 'A Partially Observed Markov Decision Process for Dynamic Pricing', *Management Science* **51**(9), 1400–1416.

Besbes, O. and Zeevi, A. (2009), 'Dynamic Pricing Without Knowing the Demand Function: Risk Bounds and Near-Optimal Algorithms'. Forthcoming in *Operations Research*.

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

Easley, D. and Kiefer, N. M. (1988), 'Controlling a Stochastic Process with Unknown Parameters', *Econometrica* **56**(5), 1045–1064.

Farias, V. F. and van Roy, B. (2009), 'Dynamic Pricing with a Prior on Market Response'. Forthcoming in *Operations Research*.

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

Keller, G. and Rady, S. (1999), 'Optimal Experimentation in a Changing Environment', *The Review of Economic Studies* **66**(3), 475–507.

Lobo, M. S. and Boyd, S. (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', *Journal of Economic Dynamics and Control* **7**(3), 331–347.

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

Robbins, H. (1951), 'Some Aspects of the Sequential Design of Experiments', *Bulletin of the American Mathematical Society* **58**, 527–535.

Rothschild, M. (1974), 'A Two-armed Bandit Theory of Market Pricing', *Journal of Economic Theory* **9**(2), 185–202.

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

Topkis, D. M. (1978), 'Minimizing a Submodular Function on a Lattice', *Operations Research* **26**, 305–321.

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