Dynamic Pricing under a General Parametric Choice Model

Josef Broder
Cornell University
jmb344@cornell.edu

Paat Rusmevichientong
Cornell University
paatrus@cornell.edu

September 3, 2010

Abstract

We consider a stylized dynamic pricing model in which a monopolist prices a product to a sequence of T customers, who independently make purchasing decisions based on the price offered according to a general parametric choice model. The parameters of the model are unknown to the seller, whose objective is to determine a pricing policy that minimizes the *regret*, which is the expected difference between the seller's revenue and the revenue of a clairvoyant seller who knows the values of the parameters in advance, and always offers the revenue-maximizing price. We show that the regret of the optimal pricing policy in this model is $\Theta(\sqrt{T})$, by establishing an $\Omega(\sqrt{T})$ lower bound on the worst-case regret under an arbitrary policy, and presenting a pricing policy based on maximum likelihood estimation whose regret is $\mathcal{O}(\sqrt{T})$ across all problem instances. Furthermore, we show that when the demand curves satisfy a "well-separated" condition, the T-period regret of the optimal policy is $\Theta(\log T)$. Numerical experiments show that our policies perform well.

1. Introduction

Consider the problem of a retailer choosing a price at which to sell a new product, with the objective of maximizing his expected revenue. If the retailer had full information about the demand at every price level, then he could determine the revenue-maximizing price for the good. However, full information about the demand curve is typically not available in practice, because the relationship between price and customer purchase probability is generally not known to the seller in advance. To address this problem, recent studies in revenue management consider *dynamic pricing* strategies, in which a seller adjusts the price of the good to gain information about the demand curve, and then exploits this information to offer a near-optimal selling price.

Regardless of the model, there are two fundamental questions that apply to virtually any dynamic pricing formulation. First, what is the value of knowing the demand curve; in other words, what is the magnitude of the revenue lost due to uncertainty about the relationship between price and demand? Secondly, how should good pricing strategies balance price experimentation (exploration) and best-guess optimal pricing (exploitation)? The answers to both of these questions depends intrinsically on the nature of the demand uncertainty facing the seller.

To investigate these questions, we consider dynamic pricing under a general parametric model of customer behavior. We measure the performance of a pricing strategy in this model in terms of the *regret*: the difference between the expected revenue gained by the pricing strategy, and the revenue gained by an omniscient strategy that has full information about the demand curve in advance. We classify the order of the regret of the optimal pricing policy under two scenarios: the fully general case, and a special case in which the parametric family of demand curves satisfies a "well-separated" condition.

By analyzing the performance of pricing policies under these two scenarios, we derive a number of insights into the above questions. We demonstrate that, in the general case, a parametric family of demand curves may include an "uninformative" price, whose presence makes dynamic pricing difficult. We demonstrate this difficulty by showing that the worst-case regret for any pricing policy in the presence of uninformative prices must be large. On the other hand, when the demand curves satisfy a well-separated condition that precludes the possibility of an uninformative price, we demonstrate the effectiveness of a "greedy" policy that simultaneously explores the demand curve and exploits at the best-guess optimal price. Intuitively, when the demand curves are "well-separated," a seller can learn from customer responses *at every price level*, making simultaneous exploration and exploitation possible, and leading to small regret.

We quantify the magnitude of the revenue lost due to demand uncertainty by providing a complete regret profile of dynamic pricing under a general parametric model, demonstrating a significant difference in the magnitude of the regret between the general and the well-separated cases. We give a detailed outline of our contributions and their organization in Section 1.3.

1.1 The Model

We assume that customers arrive in discrete time steps. For each $t \geq 1$, when the t^{th} customer arrives, he is quoted a price by the seller, and then decides whether to purchase the good at that price based on his willingness-to-pay V_t. We assume that $\{V_t : t \geq 1\}$ are independent and identically distributed random variables whose common distribution function belongs to some family parameterized by $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^n$. Let $d(\,\cdot\,;\mathbf{z}) : \mathbb{R}_+ \to \mathbb{R}_+$ denote the complementary cumulative distribution function of V_t, that is, for all $p \geq 0$,

$$d(p\,;\mathbf{z}) = \Pr_{\mathbf{z}}\{V_t \geq p\} \ . \tag{1}$$

We assume that each customer purchases the product if and only if his willingness-to-pay is at least as large as the product price. Thus, we will also refer to $d(\,\cdot\,;\mathbf{z})$ as the demand curve because it determines the probability that the customer will purchase a product at a given price. For any $p \geq 0$, the expected revenue $r(p\,;\mathbf{z})$ under the price p is given by

$$r(p\,;\mathbf{z}) = p\,d(p\,;\mathbf{z})\ . \tag{2}$$

We will restrict our attention to families of demand curves for which the corresponding revenue function $r(\,\cdot\,;\mathbf{z})$ has a unique maximum.

We will consider a *problem class* $\mathcal{C}$ to be a tuple $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$, where $\mathcal{Z} \subset \mathbb{R}^n$ is a compact and convex parameter set, $\mathcal{P} = [p_{min}, p_{max}]$ is a closed pricing interval with $p_{min} \geq 0$, and $d : \mathcal{P} \times \mathcal{Z} \to [0,1]$ is a smooth parametric family of demand curves such that $p \mapsto d(p;\mathbf{z})$ is non-increasing for each $\mathbf{z} \in \mathcal{Z}$. Finally, we assume that $p^*(\mathbf{z}) \in \mathcal{P}$ for all $\mathbf{z} \in \mathcal{Z}$.

For any $t \geq 1$, we denote by $\mathbf{y}_t = (y_1, \ldots, y_t) \in \{0,1\}^t$ a history of customer purchasing decisions, where $y_\ell = 1$ if the ℓ^{th} customer decided to purchase the product, and $y_\ell = 0$ otherwise. A *pricing policy* $\psi = (\psi_1, \psi_2, \ldots)$ is a sequence of functions such that $\psi_t : \{0,1\}^{t-1} \to \mathcal{P}$ sets the price in period t based on the observed purchasing decisions in the preceding $t-1$ periods. To model the relationship between a pricing policy ψ and customer behavior, we consider the distribution $Q_t^{\psi,\mathbf{z}}$ on t-step customer response histories induced by the policy ψ, which we define as follows. For any policy ψ and $\mathbf{z} \in \mathcal{Z}$, let $Q_t^{\psi,\mathbf{z}} : \{0,1\}^t \to [0,1]$ denote the probability distribution of the customer responses $\mathbf{Y}_t = (Y_1, \ldots, Y_t)$ in the first t periods when the policy ψ is used and the underlying parameter is $\mathbf{z}$; that is, for all $\mathbf{y}_t = (y_1, \ldots, y_t) \in \{0,1\}^t$,

$$Q_t^{\psi,\mathbf{z}}(\mathbf{y}_t) = \prod_{\ell=1}^{t} d(p_\ell;\mathbf{z})^{y_\ell}(1 - d(p_\ell;\mathbf{z}))^{1-y_\ell}\ , \tag{3}$$

where $p_\ell = \psi_\ell(\mathbf{y}_{\ell-1})$ denotes the price in period ℓ under the policy ψ. It will also be convenient to consider the distribution on customer responses to a sequence of fixed prices $\mathbf{p} = (p_1, \ldots, p_k) \in \mathcal{P}^k$, rather than the prices set by a pricing policy. We represent these distributions by

$$Q^{\mathbf{p},\mathbf{z}}(\mathbf{y}) = \prod_{\ell=1}^{k} d(p_\ell;\mathbf{z})^{y_\ell}(1 - d(p_\ell;\mathbf{z}))^{1-y_\ell},$$

where $\mathbf{y} \in \{0,1\}^k$, and p_ℓ denotes the ℓ^{th} component of the price vector $\mathbf{p} \in \mathcal{P}^k$.

Finally, we formalize the performance measure used to evaluate pricing policies. For a problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$, a parameter $\mathbf{z} \in \mathcal{Z}$, a policy ψ setting prices in $\mathcal{P}$, and a time horizon $T \geq 1$, the T-period cumulative regret under ψ is defined to be

$$\text{Regret}(\mathbf{z}, \mathcal{C}, T, \psi) = \sum_{t=1}^{T} \mathbb{E}_{\mathbf{z}}\left[r\left(p^*(\mathbf{z});\mathbf{z}\right) - r(P_t;\mathbf{z})\right]\ ,$$

where $P_1, P_2, \ldots$ denotes the sequence of prices under the policy ψ, and $\mathbb{E}_{\mathbf{z}}[\,\cdot\,]$ denotes the expectation when the underlying parameter vector of the willingness-to-pay distribution is $\mathbf{z}$. We note that when the parameter $\mathbf{z}$ is known, minimizing the T-period cumulative regret is equivalent to maximizing the total expected reward over T periods.

As a convention, we will denote vectors in bold, and scalars in regular font. A random variable is denoted by an uppercase letter while its realized values are denoted in lowercase. We denote by $\mathbb{R}_+$ the set of non-negative real numbers, while $\mathbb{R}_{++}$ denotes the set of positive numbers. We use $\|\cdot\|$ to denote the Euclidean norm, and for any set $S \subset \mathbb{R}^n$ and any element $y \in \mathbb{R}^n$, we define $S - y = \{x - y : x \in S\}$. We use $\log(\,\cdot\,)$ to denote the natural logarithm. For any symmetric matrix $\mathbf{A}$, let $\lambda_{\min}(\mathbf{A})$ denote its smallest eigenvalue.

Before proceeding with a review of the relevant literature, we note several assumptions about the retail environment implicit in our model. We assume that the seller is a monopolist offering an unlimited supply of a nonperishable single product, with no marginal cost of production. We also assume that the seller has the ability to adjust prices and receive feedback in real time, at the level of individual customers. Although quite stylized, this model allow us to conduct a simple and tractable analysis of demand learning under parametric uncertainty, and clearly illustrate some of the difficulties facing a seller in such a scenario. Moreover, these assumptions have been adopted by previous works (e.g., Cope, 2006; Kleinberg and Leighton, 2003; Carvalho and Puterman, 2005; Besbes and Zeevi, 2009), and provide a convenient framework in which to study dynamic pricing. We now proceed to place our paper in context with a review of the existing literature.

1.2 Literature Review

At a high level, many recent studies in the dynamic pricing literature focus on two natural questions. First, what are the qualitative obstacles to pricing well under demand uncertainty, and secondly, how should one design pricing strategies to overcome these obstacles? Many recent works in dynamic pricing investigate these questions by considering numerical evaluations of heuristic policies, focusing mainly on the second question posed above. Carvalho and Puterman (2005) consider a dynamic pricing formulation in which the demand has a logistic distribution with two unknown parameters. The authors perform a numerical evaluation of several heuristic strategies, and demonstrate that a "one-step lookahead" policy, which sacrifices immediate revenue to compute a better estimate of the unknown demand parameters, outperforms a myopic policy. Lobo and Boyd (2003) consider a linear demand model with Gaussian noise, and investigate through numerical experiments a "price-dithering" policy, which adds a random perturbation to the myopically optimal price. Bertsimas and Perakis (2003) consider a similar demand model, and show through numerical experiments that approximate dynamic programming policies that balance immediate revenue rewards with long-term learning can outperform a myopic policy.

The above works provide empirical evidence that, in a variety of settings, pricing policies that perform some sort of active exploration will outperform myopically greedy policies, indicating that there is some intrinsic value to price experimentation. Several other recent papers conduct a more theoretical investigation into the value of price experimentation. In Besbes and Zeevi (2009), the authors consider demand learning under an uncapacitated Bernoulli demand model, in which the seller knows the initial demand curve. At some point in time unknown to the seller, the demand curve switches to a different (but known in advance) function of the price. The authors show that when the two demand curves satisfy a well-separated condition, a myopically greedy policy is optimal. Additionally, they show that when the demand curves intersect, corresponding to the presence of an uninformative price, then the magnitude of the worst-case regret is larger, and exhibit an optimal policy that performs some forced exploration. Our work in this paper is thematically related to Besbes and Zeevi (2009), in that we conduct a similar analysis of the worst-case regret under a well-separated versus intersecting demand model, and in that we consider myopic versus forced exploration policies. One may view our work as complementary to Besbes and Zeevi (2009), in that we consider demand learning in a stationary, parameter learning framework, while they consider a similar learning problem under a non-stationary, two-hypothesis testing setting.

A second related paper by the same authors is Besbes and Zeevi (2008). Here, the authors consider demand learning in a general parametric (as well as non-parametric) setting, and present policies based on maximum likelihood estimation. They suggest that the structure and performance of a rate-optimal pricing policy should be different in the general versus the well-separated case, but they provide the same lower bound on the performance measure for both cases. We complement the theme of their work by exhibiting a dynamic pricing formulation in which the regret profiles between the two cases are entirely different. Specifically, we prove in Theorem 3.1 that in the general case, the worst case regret under an arbitrary policy must be at least $\Omega(\sqrt{T})$. On the other hand, in the well-separated case, there is a policy whose regret is at most $\mathcal{O}(\log T)$ in all problem instances (Theorem 4.8). Aside from these thematic similarities, several crucial features differentiate this work from ours, including the presence of a known, finite time horizon, the presence of a known capacity constraint, and a performance measure that is parameterized by initial capacity and demand rate, rather than the time horizon. While we present pricing policies with similar structure to those presented in Besbes and Zeevi (2008), the aforementioned differences make direct comparisons difficult, and lead to a significantly different analysis.

Other recent related results include Lim and Shanthikumar (2007), which considers dynamic pricing strategies that are robust in the face of model uncertainty, using the classical pricing framework of Gallego and van Ryzin (1994). Harrison et al. (2010) considers a stationary, two-hypothesis dynamic pricing problem from a Bayesian standpoint. As in Besbes and Zeevi (2009) and this paper, they consider problem instances in which there exists an "uninformative price," and demonstrate that a myopically greedy Bayesian policy can perform poorly in the presence of these uninformative prices. The authors then analyze variants of the myopic Bayesian policy that perform active exploration, and show that the revenue loss of these policies does not grow with the time horizon. den Boer and Zwart (2010) consider dynamic pricing in a two-parameter linear demand model, with Gaussian noise. The authors consider a myopically greedy policy based on least-squares parameter estimation, and show that with positive probability, the prices generated by the myopic policy do not converge to the optimal price. The authors then propose an alternative to the myopically greedy policy, called "controlled variance pricing," that maintains a "taboo" interval around the average of the prices offered up to time t, and then offers the best-guess optimal price outside of this interval. The size of the interval is carefully controlled to balance immediate revenue maximization with long-term demand learning, resulting in a policy that is essentially optimal, up to logarithmic factors.

Finally, we note that our pricing problem can be viewed as a special case of a general stochastic optimization problem, in which one wishes to iteratively approximate the minimizer of an unknown function, based only on noisy evaluation of the function at points inside a (usually uncountable) feasible set. A full review of the literature on this topic is beyond the scope of this paper; however, several notable references from the stochastic approximations literature include Kiefer and Wolfowitz (1967), Fabian (1967), and more recently, Broadie et al. (2009) and Cope (2009), which examine the convergence properties of stochastic gradient-descent type schemes. Another standard approach is to apply the classical multi-armed bandit algorithm (Lai and Robbins, 1985 and Auer et al., 2002) to the general stochastic optimization setting via a discretization approach; see, for example, Agrawal (1995) and Auer et al. (2007), and Kleinberg and Leighton (2003) for an application of these techniques in the context of dynamic pricing. As a key distinction, we note that both of the aforementioned techniques are *non-parametric*, and thus the parametric, maximum-likelihood-based policies presented in this paper are significantly different in both their structure and analysis.

We now proceed with a summary of our main contributions and organization.

1.3 Contributions and Organization

One of the main contributions of our work is a complete regret profile for the dynamic pricing problem under a general parametric choice model. In Section 3.1, we prove in Theorem 3.1 that in the general case, the regret of an arbitrary pricing policy is $\Omega(\sqrt{T})$, by exploiting the presence of "uninformative prices," which force a tradeoff between reducing uncertainty about the parameters of the demand curve and exploiting the best-guess optimal price.[1] In Section 3.2, we present a pricing policy based on maximum-likelihood estimation whose regret is $\mathcal{O}(\sqrt{T})$ across all problem instances (Theorem 3.6).

In Section 4, we consider dynamic pricing when the family of demand curves satisfies a "well-separated" condition, which precludes the presence of uninformative prices. We show that in this scenario, the regret of the optimal policy is $\Theta(\log T)$. In Section 4.1, we establish a regret lower bound of $\Omega(\log T)$ for all policies (Theorem 4.1), based on a Cramér-Rao-type inequality. We also describe a pricing policy based on maximum-likelihood estimates (MLE) that achieves a matching $\mathcal{O}(\log T)$ upper bound (Theorem 4.8). The key observation is that in the well-separated case, demand learning is easier, in that a pricing policy can learn about the parameters of the demand curve from customer responses to any price.

As a by product of our analysis, we also provide a novel large deviation inequality and bound on mean squared errors for a maximum-likelihood estimator based on samples that are dependent and not identically distributed (Theorem 4.7). The proof techniques used here are of independent interest because they can be extended to other MLE-based online learning strategies.

2. Assumptions and Examples

Recall that a *problem class* $\mathcal{C}$ is a tuple $(\mathcal{P}, \mathcal{Z}, d)$, where $\mathcal{P} = [p_{min}, p_{max}] \subset \mathbb{R}_+$ is a feasible pricing interval, $\mathcal{Z} \subset \mathbb{R}^n$ is a compact and convex feasible parameter set, and $d : \mathcal{P} \times \mathcal{Z} \to [0,1]$ is a parametric family of smooth demand functions such that $p \mapsto d(p; \mathbf{z})$ is non-increasing for each $\mathbf{z} \in \mathcal{Z}$. Throughout the paper, we restrict our attention to problem classes $\mathcal{C}$ satisfying the following basic assumptions.

Assumption 1 (Basic Assumptions)**.** There exists positive constants d_{min}, d_{max}, L, and c_r such that

(a) $0 < d_{min} \leq d(p; \mathbf{z}) \leq d_{max} < 1$ for all $p \in \mathcal{P}$ and $\mathbf{z} \in \mathcal{Z}$.

(b) The revenue function $p \mapsto r(p; \mathbf{z})$ has a unique maximizer $p^*(\mathbf{z}) \in \mathcal{P}$.

(c) The function $\mathbf{z} \mapsto p^*(\mathbf{z})$ is L-Lipschitz, that is, $|p^*(\mathbf{z}) - p^*(\bar{\mathbf{z}})| \leq L \left\| \mathbf{z} - \bar{\mathbf{z}} \right\|$ for all $\mathbf{z}, \bar{\mathbf{z}} \in \mathcal{Z}$.

(d) The revenue function $p \mapsto r(p; \mathbf{z})$ is twice differentiable with $\sup_{p \in \mathcal{P}, \mathbf{z} \in \mathcal{Z}} |r''(p; \mathbf{z})| \leq c_r$.

Under Assumption 1(a), the demand is bounded away from zero and one on the pricing interval;

[1] We use the notation $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ to represent upper and lower bounds, respectively, on the performance measure of interest (see Knuth, 1997 for more details).

that is, we will not offer prices at which customers will either purchase or decline to purchase with probability one. Assumption 1(b) is self-explanatory, and Assumption 1(c) says that if we vary the parameter $\mathbf{z}$ by a small amount, then the optimal price $p^*(\mathbf{z})$ will not vary too much. Assumption 1(d) imposes a smoothness condition on the demand curve $p \mapsto d(p; \mathbf{z})$.

In addition to these structural assumptions about the demand curve, we will also impose the following statistical assumption about the family of distributions $\{Q^{\mathbf{p},\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$.

Assumption 2 (Statistical Assumption)**.** There exists a vector of exploration prices $\bar{\mathbf{p}} \in \mathcal{P}^k$ such that the family of distributions $\{Q^{\bar{\mathbf{p}},\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ is identifiable, that is, $Q^{\bar{\mathbf{p}},\mathbf{z}}(\,\cdot\,) \neq Q^{\bar{\mathbf{p}},\bar{\mathbf{z}}}(\,\cdot\,)$ whenever $\mathbf{z} \neq \bar{\mathbf{z}}$. Moreover, there exists a constant $c_f > 0$ depending only on the problem class $\mathcal{C}$ and $\bar{\mathbf{p}}$ such that $\lambda_{min}\{\mathbf{I}(\bar{\mathbf{p}}, \mathbf{z})\} \geq c_f$ for all $\mathbf{z} \in \mathcal{Z}$, where $\mathbf{I}(\bar{\mathbf{p}}, \mathbf{z})$ denotes the *Fisher information matrix* given by

$$\left[\mathbf{I}(\bar{\mathbf{p}}, \mathbf{z})\right]_{i,j} = \mathbb{E}_{\mathbf{z}}\left[-\frac{\partial^2}{\partial z_i \partial z_j} \log Q^{\bar{\mathbf{p}},\mathbf{z}}(\mathbf{Y})\right] = \sum_{k=1}^{n} \frac{\left\{\frac{\partial}{\partial z_i} d(\bar{p}_k, \mathbf{z})\right\} \times \left\{\frac{\partial}{\partial z_j} d(\bar{p}_k, \mathbf{z})\right\}}{d(\bar{p}_k, \mathbf{z})(1 - d(\bar{p}_k, \mathbf{z}))} \ .$$

Assumption 2 is a standard assumption, which guarantees that we can estimate the demand parameter based on the purchase observations at the exploration prices $\bar{\mathbf{p}}$ (see, for example, Besbes and Zeevi, 2008). As shown in the following examples, Assumptions 1 and 2 encompass many families of parametric demand curves (see Talluri and van Ryzin, 2004 for additional examples).

Example 2.1 (Logit Demand)**.** Let $\mathcal{P} = [1/2, 2] \subset \mathbb{R}$, $\mathcal{Z} = [1, 2] \times [-1, 1] \subset \mathbb{R}^2$ and let

$$d(p, \mathbf{z}) = \frac{e^{-z_1 p - z_2}}{1 + e^{-z_1 p - z_2}}$$

be the family of logit demand curves. It is straightforward to check that $(\mathcal{P}, \mathcal{Z}, d)$ satisfies the conditions stated in Assumption 1 with $d_{\min} = e^{-5}/(1+e^{-5})$, $d_{\max} = e^{1/2}/(1+e^{1/2})$, $L = 2+\log(2)$, and $c_r = 2e$. It is also straightforward to check that for any $\bar{\mathbf{p}} = (\bar{p}_1, \bar{p}_2) \in \mathcal{P}^2$ with $\bar{p}_1 \neq \bar{p}_2$, the associated family $\{Q^{\bar{\mathbf{p}},\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ is identifiable. Moreover, for any $p \in \mathbb{R}_+$ and $\mathbf{z} \in \mathcal{Z}$, we have that

$$\frac{\partial}{\partial z_1} d(p, \mathbf{z}) = -p d(p, \mathbf{z})(1 - d(p, \mathbf{z})) \quad \text{and} \quad \frac{\partial}{\partial z_2} d(p, \mathbf{z}) = -d(p, \mathbf{z})(1 - d(p, \mathbf{z})) \ ,$$

which implies that the Fisher information matrix is given by

$$\mathbf{I}(\bar{\mathbf{p}}, \mathbf{z}) = d(\bar{p}_1, \mathbf{z})(1 - d(\bar{p}_1, \mathbf{z})) \begin{pmatrix} \bar{p}_1^2 & \bar{p}_1 \\ \bar{p}_1 & 1 \end{pmatrix} + d(\bar{p}_2, \mathbf{z})(1 - d(\bar{p}_2, \mathbf{z})) \begin{pmatrix} \bar{p}_2^2 & \bar{p}_2 \\ \bar{p}_2 & 1 \end{pmatrix}$$

By applying the trace-determinant formula, we can show that for all $\mathbf{z} \in \mathcal{Z}$,

$$\lambda_{min}\{\mathbf{I}(\bar{\mathbf{p}}, \mathbf{z})\} \geq \frac{(\bar{p}_1 - \bar{p}_2)^2}{\bar{p}_1^2 + \bar{p}_2^2 + 2} \cdot d_{min}^2 (1 - d_{max})^2 > 0$$

Example 2.2 (Linear Demand)**.** Let $\mathcal{P} = [1/3, 1/2]$, let $\mathcal{Z} = [2/3, 3/4] \times [3/4, 1]$, and let

$$d(p; \mathbf{z}) = z_1 - z_2 p$$

be a linear demand family. Then it is straightforward to check that this family satisfies Assumption 1 with $d_{\min} = 1/6$, $d_{\max} = 1/2$, $L = 2$, and $c_r = 2$. Moreover, for any $\bar{\mathbf{p}} = (\bar{p}_1, \bar{p}_2) \in \mathcal{P}^2$ with $\bar{p}_1 \neq \bar{p}_2$, the associated family $\{Q^{\bar{\mathbf{p}},\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ is identifiable. A similar computation shows that the Fisher information matrix is given by

$$\mathbf{I}(\bar{\mathbf{p}}, \mathbf{z}) = \frac{1}{d(\bar{p}_1, \mathbf{z})(1 - d(\bar{p}_1, \mathbf{z}))} \begin{pmatrix} 1 & \bar{p}_1 \\ \bar{p}_1 & \bar{p}_1^2 \end{pmatrix} + \frac{1}{d(\bar{p}_2, \mathbf{z})(1 - d(\bar{p}_2, \mathbf{z}))} \begin{pmatrix} 1 & \bar{p}_2 \\ \bar{p}_2 & \bar{p}_2^2 \end{pmatrix} ,$$

and using the same argument as above, we can show that

$$\lambda_{min}\{\mathbf{I}(\bar{\mathbf{p}}, \mathbf{z})\} \geq \frac{(\bar{p}_1 - \bar{p}_2)^2}{\bar{p}_1^2 + \bar{p}_2^2 + 2} \cdot \frac{1}{d_{max}^2 (1 - d_{min})^2} > 0 .$$

Example 2.3 (Exponential Demand)**.** Let $\mathcal{P} = [1/2, 1]$, let $\mathcal{Z} = [1, 2] \times [0, 1]$, and let

$$d(p; \mathbf{z}) = e^{-z_1 p - z_2}$$

be an exponential demand family. Then by the same techniques used in Examples 2.1 and 2.2, one may check that this problem class satisfies Assumptions 1 and 2, and that the associated family $\{Q^{\bar{\mathbf{p}},\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ is identifiable for any $\bar{\mathbf{p}} = (p_1, p_2) \in \mathcal{P}^2$ for which $p_1 \neq p_2$.

We now discuss an important observation that will motivate the design of pricing policies in our model. Suppose that the unknown model parameter vector is $\mathbf{z}$, and let $\hat{\mathbf{z}}$ denote some estimate of $\mathbf{z}$. We might consider pricing the product at $p^*(\hat{\mathbf{z}})$, which is optimal with respect to our estimate. When $\hat{\mathbf{z}}$ is close to the true parameter vector, we would expect that $p^*(\hat{\mathbf{z}})$ yields a near optimal revenue. We make this intuition precise in the following corollary, which establishes an upper bound on the loss in revenue from inaccurate estimation.

Corollary 2.4 (Revenue Loss from Inaccurate Estimation)**.** *For any problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ satisfying Assumption 1 and for any $\mathbf{z}, \hat{\mathbf{z}} \in \mathcal{Z}$,*

$$r(p^*(\mathbf{z}); \mathbf{z}) - r(p^*(\hat{\mathbf{z}}); \mathbf{z}) \ \leq \ c_r \, L^2 \, \|\mathbf{z} - \hat{\mathbf{z}}\|^2 \ .$$

Proof. First we will show that as a consequence of Assumption 1(b) and Assumption 1(d), we have that for $\mathbf{z} \in \mathcal{Z}$ and $p \in \mathcal{P}$,

$$0 \leq r(p^*(\mathbf{z}); \mathbf{z}) - r(p; \mathbf{z}) \leq c_r \, (p^*(\mathbf{z}) - p)^2 \ .$$

The result then follows from Assumption 1(c) (the Lipschitz continuity of the optimal price).

We will establish the quadratic inequality for $p > p^*(\mathbf{z})$. The same argument applies to the case where $p < p^*(\mathbf{z})$. For any $u \in \mathbb{R}_+$, let $r'(u;\mathbf{z})$ and $r''(u;\mathbf{z})$ denote the first and second derivatives of the revenue function at u, respectively. Since $r'(p^*(\mathbf{z});\mathbf{z}) = 0$, it follows that

$$\begin{aligned} |r(p^*(\mathbf{z});\mathbf{z}) - r(p;\mathbf{z})| &= \left| \int_{p^*(\mathbf{z})}^{p} \int_{p^*(\mathbf{z})}^{t} r''(u;\mathbf{z})\, du\, dt \right| \\ &\leq \sup_{u\in\mathcal{P}} \left| r''(u;\mathbf{z}) \right| \int_{p^*(\mathbf{z})}^{p} \int_{p^*(\mathbf{z})}^{t} du\, dt \;=\; \frac{1}{2} \sup_{u\in\mathcal{P}} \left| r''(u;\mathbf{z}) \right| (p^*(\mathbf{z}) - p)^2 \\ &\leq c_r (p^*(\mathbf{z}) - p)^2 \end{aligned}$$

□

Corollary 2.4 suggests a method for constructing a pricing policy with low regret. We construct an estimate of the underlying parameter based on the observed purchase history, then offer the greedy optimal price according to this estimate. If our estimate has a small mean square error, then we expect that the loss in revenue should also be small. However, the variability of our estimates depends on the past prices offered. As we will see, there is a nontrivial tradeoff between pricing to form a good estimate (exploration) and pricing near the greedy optimal (exploitation), and the optimal balance between these two will be quite different depending on the nature of the demand uncertainty facing the seller.

3. The General Case

In this section, we consider dynamic pricing under the general parametric model satisfying Assumptions 1 and 2. In Section 3.1, we show that the worst-case regret of any pricing policy must be at least $\Omega(\sqrt{T})$, by constructing a problem class with an "uninformative price" that impedes demand learning. Then, in Section 3.2, we describe a pricing policy based on maximum likelihood estimation whose regret is $\mathcal{O}(\sqrt{T})$ across all problem instances, thus establishing that the order of regret for the optimal pricing policy in the general case is $\Theta(\sqrt{T})$.

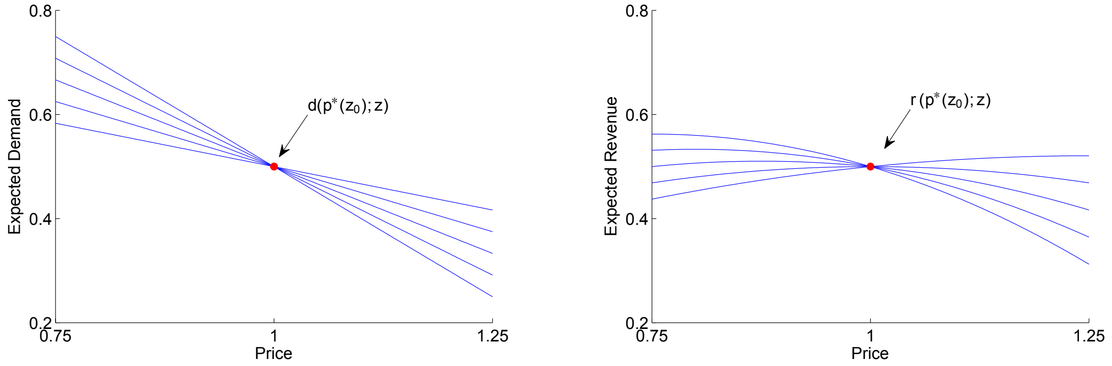
3.1 A Lower Bound for the General Case

In this section, we establish a lower bound on the T-period cumulative regret for the general case. The main result is stated in the following theorem.

Theorem 3.1 (General Regret Lower Bound)**.** *Define a problem class $\mathcal{C}_{\mathsf{GenLB}} = (\mathcal{P}, \mathcal{Z}, d)$ by letting $\mathcal{P} = [3/4, 5/4]$, $\mathcal{Z} = [1/3, 1]$, and $d(p;z) = 1/2 + z - zp$. Then for any policy ψ setting prices in $\mathcal{P}$, and any $T \geq 2$, there exists a parameter $z \in \mathcal{Z}$ such that*

$$\text{Regret}(z, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) \geq \frac{\sqrt{T}}{48^3}\ .$$

Using the same proof technique as in Example 2.2, one can show that the problem class $\mathcal{C}_{\mathsf{GenLB}}$ satisfies Assumptions 1 and 2, with $d_{min} = 1/4, d_{max} = 3/4, p^*(z) = (1+2z)/(4z), L = 3$, and $c_r = 2$. Before we proceed to the proof of Theorem 3.1, let us discuss the intuition underlying our arguments. Figure 1(a) shows examples of demand curves in the family given by $\mathcal{C}_{\mathsf{GenLB}}$. Note that for all $z \in \mathcal{Z}$, $d(1; z) = 1/2$, and thus all demand curves in this family intersect at common price $p = 1$. Note also that this price is the optimal price for some demand curve in this family, that is, $p^*(z_0) = 1$ for $z_0 = 1/2$ (see Figure 1(b) for examples of the revenue curves). Since the demand is the same at $p^*(z_0)$ regardless of the underlying parameter, the price $p^*(z_0)$ is "uninformative," in that no policy can gain information about the value of the parameter while pricing at $p^*(z_0)$. To establish Theorem 3.1, we show that uninformative prices lead to a tension between demand learning (exploration) and best-guess optimal pricing (exploitation), which forces the worst-case regret of any policy to be $\Omega(\sqrt{T})$. This tension is made precise in two lemmas. We show in Lemma 3.3 that for a policy to reduce its uncertainty about the unknown demand parameter, it must necessarily set prices away from the uninformative price $p^*(z_0)$, and thus incur large regret when the underlying parameter is z_0. Then, in Lemma 3.4, we show that any policy that does not reduce its uncertainty about the demand parameter z must also incur a cost in regret.



(a) expected demand

(b) expected revenue

Figure 1: Family of linear demand and revenue curves under $\mathcal{C}_{\mathsf{GenLB}}$ for $z \in \{1/3, 1/2, 2/3, 5/6, 1\}$. For $z = 1/2$, the optimal price is $p^*(1/2) = 1$, which is also the common intersection points for all demand curves in this family.

To give precise statements of Lemmas 3.3 and 3.4, we will need to quantify the heuristic notion of "uncertainty" about the unknown demand parameter. In our analysis, we will use a convenient quantitative measure of uncertainty, known as the *KL divergence*.

Definition 3.2 (Definition 2.26 in Cover and Thomas, 1999)**.** For any probability measures Q_0 and

Q_1 on a discrete sample space $\mathcal{Y}$, the *KL divergence* of Q_0 and Q_1 is

$$\mathcal{K}\left(Q_0; Q_1\right) = \sum_{y \in \mathcal{Y}} Q_0(y) \log\left(\frac{Q_0(y)}{Q_1(y)}\right) .$$

Intuitively, the KL divergence is a measure of distinguishability between two distributions; if the KL divergence between Q_0 and Q_1 is large, then Q_0 and Q_1 are easily distinguishable, and if $\mathcal{K}(Q_0; Q_1)$ is small, then Q_0 and Q_1 are difficult to distinguish. Thus, we say that a pricing policy ψ has a large degree of certainty that the true underlying demand parameter is z_0, rather than some counterfactual parameter z, if the quantity $\mathcal{K}(Q_t^{\psi,z_0}; Q_t^{\psi,z})$ is large.

With this interpretation of the KL divergence, we now state Lemma 3.3. This lemma establishes that reducing uncertainty about the underlying parameter is costly, by establishing a lower bound on the regret incurred by an arbitrary pricing policy in terms of the KL divergence.

Lemma 3.3 (Learning is Costly)**.** *For any $z \in \mathcal{Z}$, $t \geq 1$, and any policy ψ setting prices in $\mathcal{P}$,*

$$\mathcal{K}\left(Q_t^{\psi,z_0}; Q_t^{\psi,z}\right) \leq \frac{9}{16}\left(z_0 - z\right)^2 \operatorname{Regret}\left(z_0, \mathcal{C}_{\mathsf{GenLB}}, t, \psi\right) \ ,$$

where $z_0 = 1/2$.

The proof of Lemma 3.4 is deferred to Appendix A.1, but here we give a high level description of the argument. Suppose the underlying demand parameter is z_0, and suppose a pricing policy ψ has the goal of reducing its uncertainty about whether the underlying demand parameter is in fact z_0, as opposed to some other value z. We may restate this goal of "reducing uncertainty" in terms of the KL divergence, by saying that the policy ψ wishes to offer a sequence of prices such that the KL divergence between the induced distributions Q_t^{ψ,z_0} and $Q_t^{\psi,z}$ of customer responses is large. To accomplish this, ψ must offer prices at which the customer purchase probability will be significantly different under z_0 versus z; however, for all prices in a small neighborhood of the uninformative price $p^*(z_0)$, the probability of a customer purchase is virtually the same under z_0 and z. Thus, to distinguish the two cases (that is, increase the KL divergence $\mathcal{K}(Q_t^{\psi,z_0}; Q_t^{\psi,z})$), the policy ψ must offer prices away from $p^*(z_0)$, and thus incur large regret $\operatorname{Regret}\left(z_0, \mathcal{C}_{\mathsf{GenLB}}, t, \psi\right)$ when the underlying parameter is in fact z_0.

We have now established in Lemma 3.3 that reducing uncertainty about the underlying demand curve is costly. However, this result alone is not enough to prove a lower bound on the regret. To establish the desired lower bound on regret, we need a complementary result, showing that any pricing policy that does not decrease its uncertainty about the demand curve must also incur a cost in regret. We establish this complement to Lemma 3.3 in the following lemma.

Lemma 3.4 (Uncertainty is Costly). *Let ψ be any pricing policy setting prices in $\mathcal{P}$ Then, for any $T \geq 2$ and for demand parameters $z_0 = 1/2$ and $z_1 = z_0 + \frac{1}{4}T^{-1/4}$, we have*

$$\text{Regret}(z_0, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) + \text{Regret}(z_1, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) \geq \frac{\sqrt{T}}{12(48^2)} e^{-\mathcal{K}(Q_T^{\psi,z_0}; Q_T^{\psi,z_1})} \ .$$

The intuition for Lemma 3.4 is the following. Let us choose the special parameter $z_0 = 1/2$ such that the corresponding optimal price $p^*(z_0)$ is the uninformative price, and let us choose a second demand parameter $z_1 = z_0 + \frac{1}{4}T^{-1/4}$. The parameters z_0 and z_1 are chosen so that the optimal prices $p^*(z_0)$ and $p^*(z_1)$ are not too close to each other; in other words, z_0 and z_1 are far enough apart (with respect to the time horizon T) such that a near-optimal pricing decision when the demand parameter is z_0 will be sub-optimal when the demand parameter is z_1, and vice versa. Thus, for a pricing policy ψ to price well under both z_0 and z_1, it must be able to distinguish which of the two is the true demand parameter, based on observed responses to the past prices offered. Consequently, if ψ cannot distinguish between the two cases z_0 and z_1 based on past prices offered (that is, the KL divergence $\mathcal{K}(Q_t^{\psi,z_0}; Q_t^{\psi,z_1})$ is small), then the worst-case regret of ψ must necessarily be large, as seen in the inequality of Lemma 3.4.

The proof of Lemma 3.4 follows from standard results on the minimum error probability of a two-hypothesis test, and we give a fully detailed proof in Appendix A.2. Equipped with Lemmas 3.3 and 3.4, we can immediately deduce the main result.

Proof of Theorem 3.1. Since $\text{Regret}(z_1, \mathcal{C}_{\mathsf{GenLB}}, T, \psi)$ is non-negative, and since $z_1 = z_0 + \frac{1}{4}T^{-1/4}$ by definition, it follows from Lemma 3.3 and the choice of z_1 that

$$\text{Regret}(z_0, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) + \text{Regret}(z_1, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) \geq \frac{\sqrt{T}}{9} \mathcal{K}\left(Q_T^{\psi,z_0}; Q_T^{\psi,z_1}\right) \ .$$

Adding this inequality to the result of Lemma 3.4, and using the fact that the KL divergence is non-negative, we have

$$\begin{aligned}
&2\left\{\text{Regret}(z_0, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) + \text{Regret}(z_1, \mathcal{C}_{\mathsf{GenLB}}, T, \psi)\right\} \\
&\geq \frac{\sqrt{T}}{9} \mathcal{K}\left(Q_T^{\psi,z_0}; Q_T^{\psi,z_1}\right) + \frac{\sqrt{T}}{12(48^2)} e^{-\mathcal{K}(Q_T^{\psi,z_0}; Q_T^{\psi,z_1})} \\
&\geq \frac{\sqrt{T}}{12(48^2)} \cdot \left\{\mathcal{K}\left(Q_T^{\psi,z_0}; Q_T^{\psi,z_1}\right) + e^{-\mathcal{K}(Q_T^{\psi,z_0}; Q_T^{\psi,z_1})}\right\} \ \geq \ \frac{\sqrt{T}}{12(48^2)}.
\end{aligned}$$

To see the last inequality, note that $\mathcal{K}\left(Q_T^{\psi,z_0}; Q_T^{\psi,z_1}\right) + e^{-\mathcal{K}(Q_T^{\psi,z_0}; Q_T^{\psi,z_1})} \geq 1$, since $x + e^{-x} \geq 1$ for all $x \in \mathbb{R}_+$. Thus, the tension between pricing optimally and learning the parameters of the demand curve is captured explicitly by the sum $\mathcal{K}\left(Q_T^{\psi,z_0}; Q_T^{\psi,z_1}\right) + e^{-\mathcal{K}(Q_T^{\psi,z_0}; Q_T^{\psi,z_1})}$. The first term in the sum captures the cost of learning the parameters of the demand curve, while the second term in the sum captures the cost of uncertainty. The fact that this sum cannot be driven to zero, regardless of the choice of the pricing policy, captures the tradeoff between learning and exploiting in the presence of uninformative prices. The desired result follows from the fact that

$$\max_{z \in \{z_0, z_1\}} \text{Regret}(z, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) \geq \frac{\text{Regret}(z_0, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) + \text{Regret}(z_1, \mathcal{C}_{\mathsf{GenLB}}, T, \psi)}{2} \geq \frac{\sqrt{T}}{48^3}$$

□

Remark 3.5 (Statistical Identifiability)**.** The result of Theorem 3.1 leverages the presence of an "uninformative price" $p^*(z_0)$. Note that the family of distributions $\{Q^{p^*(z_0),z} : z \in \mathcal{Z}\}$ is not identifiable, that is, one cannot uniquely identify the true value of the underlying demand parameter z from observing customer responses *to the single price* $p^*(z_0)$. However, by the arguments of Example 2.2, the family $\{Q^{\bar{\mathbf{p}},z} : z \in \mathcal{Z}\}$ *is* identifiable for any $\bar{\mathbf{p}} = (p_1, p_2) \in \mathcal{P}^2$ with $p_1 \neq p_2$, that is, one *can* uniquely identify the value of the underlying parameter from observing customer responses to two distinct prices.

Before we proceed with Section 3.2, we briefly remark on the related literature. A very general version of the result of Theorem 3.1 was previously known in the computer science literature; Kleinberg and Leighton (2003) contains eight sufficient conditions under which a one-parameter family of demand curves yields regret that is not $o(\sqrt{T})$. It is worth noting that the family constructed in Theorem 3.1 does not satisfy the sufficient conditions provided by Kleinberg and Leighton (2003); in particular, the family presented in Theorem 3.1 contains an "uninformative price," while their lower bound proof exploits alternative properties.

The techniques used in the proof of Theorem 3.1 have appeared in several recent papers. A recent work in dynamic pricing is Besbes and Zeevi (2009), which contains a related lower bound result in a non-stationary demand learning framework. Examples of these techniques in the more general online learning literature can be found in Goldenshluger and Zeevi (2008) and Goldenshluger and Zeevi (2009), which concern optimal learning in a two-armed bandit setting.

3.2 A General Matching Upper Bound

In this section, we present a pricing policy called MLE-CYCLE whose regret is $\mathcal{O}(\sqrt{T})$ across all problem instances, matching the order of the lower bound of Section 3.1. We describe the policy MLE-CYCLE in detail below, but first we describe the general intuition behind the policy.

Suppose we had access to a good estimate of the underlying demand parameter. Then this would give us a good approximation of the true demand curve, and we would be able to price near-optimally (per the result of Corollary 2.4). However, any estimate of the demand parameter will depend on customer responses to the past prices offered, and as seen in Theorem 3.1, observing responses to prices near an "uninformative price" will do little to reduce uncertainty about the demand parameter. Thus, to learn the demand curve adequately, a pricing policy should be careful to offer prices at which a good estimate of the demand parameter can be computed.

Motivated by this discussion, we present a policy MLE-CYCLE based on maximum likelihood parameter estimation. The policy MLE-CYCLE operates in cycles, and each cycle consists of an exploration phase followed by an exploitation phase. These cycles are simply a scheduling device, designed to maintain the appropriate balance between exploration and exploitation. During the exploration phase of a given cycle c, we offer the product to consecutive customers at a sequence of exploration prices $\mathbf{p} \in \mathcal{P}^k$, and then compute a maximum likelihood estimate of the underlying parameter based on the observed customer selections. The exploration prices $\mathbf{p}$ are fixed, and are chosen so that a good estimate of the demand parameter can be computed from the corresponding customer responses. Following the exploration phase of cycle c, there is an exploitation phase of c periods, during which we offer the best-guess optimal price corresponding to the current estimate of the demand parameter to c consecutive customers. Thus, the c^{th} cycle of MLE-CYCLE consists of $(k+c)$ periods: k periods in which we offer each of the k exploration prices, followed by c periods in which we offer the optimal price corresponding to our most recent estimate of the demand parameter. The cycle-based scheduling of MLE-CYCLE is carefully chosen to optimize the balance the amount of demand learning (exploration) with best-guess optimal pricing (exploitation). While we make this balance precise in the analysis of the policy, we note that the scheduling makes intuitive sense: the ratio of exploration steps to exploitation steps in MLE-CYCLE is high in the early time periods, when little is known about the demand curve, and is low in the later time periods, when the demand curve is known to a good approximation.

We now proceed with a formal description of the policy MLE-CYCLE.

We state the regret guarantee of MLE-CYCLE in the following theorem.

Theorem 3.6 (General Regret Upper Bound). *For any problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ satisfying Assumptions 1 and 2 with corresponding exploration prices $\bar{\mathbf{p}} \in \mathcal{P}^k$, there exists a constant C_1 depending only on the exploration prices $\bar{\mathbf{p}}$ and the problem class $\mathcal{C}$ such that for all $\mathbf{z} \in \mathcal{Z}$ and $T \geq 2$, the policy* MLE-CYCLE *satisfies*

$$\text{Regret}(\mathbf{z}, \mathcal{C}, T, \text{MLE-CYCLE}) \leq C_1 \sqrt{T} \ .$$

The main idea of the proof of Theorem 3.6 is the following. For a given time horizon T, it is straightforward to check that the number of cycles up to time T is $\mathcal{O}(\sqrt{T})$, and so to prove that the regret of MLE-CYCLE is $\mathcal{O}(\sqrt{T})$, it is enough to show that the regret in each cycle is $\mathcal{O}(1)$. Since each cycle consists of an exploration phase followed by an exploitation phase, it's enough to show

Policy $\mathsf{MLE\text{-}CYCLE}(\mathcal{C}, \mathbf{p})$

Inputs: A problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ and exploration prices $\bar{\mathbf{p}} = (\bar{p}_1, \ldots, \bar{p}_k) \in \mathcal{P}^k$.

Description: For each cycle $c = 1, 2, \ldots,$

- Exploration Phase (k periods): Offer the product at exploration prices $\bar{\mathbf{p}} = (\bar{p}_1, \ldots, \bar{p}_k)$ and let $\mathbf{Y}(c) = (Y_1(c), \ldots, Y_k(c))$ denote the corresponding customer selections. Let $\widehat{\mathbf{Z}}(c)$ denote the maximum likelihood estimate (MLE) based on observed customer selections during the exploration phases in the past c cycles, that is,

$$\widehat{\mathbf{Z}}(c) = \arg\max_{\mathbf{z} \in \mathcal{Z}} \prod_{s=1}^{c} Q^{\bar{\mathbf{p}}, \mathbf{z}}(\mathbf{Y}(s)) \ ,$$

  where for each $1 \leq s \leq c$, $\mathbf{Y}(s) = (Y_1(s), \ldots, Y_k(s))$ denotes the observed customer responses to the exploration prices offered in the exploration phase of cycle s.

- Exploitation Phase (c periods): Offer the greedy price $p^*\left(\widehat{\mathbf{Z}}(c)\right)$ based on the estimate $\widehat{\mathbf{Z}}(c)$.

that for an arbitrary cycle c, the regret incurred in the exploration phase is $\mathcal{O}(1)$, and the regret incurred during the exploitation phase is $\mathcal{O}(1)$.

First, to show that the regret during the exploration phase of an arbitrary cycle is $\mathcal{O}(1)$, note that during the exploration phase, $\mathsf{MLE\text{-}CYCLE}$ offers k exploration prices, and the regret incurred from offering each of these exploration prices is $\mathcal{O}(1)$, by the smoothness of the revenue function, and the compactness of the pricing interval. Thus, the total regret incurred during the exploration phase is $\mathcal{O}(1)$. Secondly, to show that the regret incurred during the exploitation phase of an arbitrary cycle is $\mathcal{O}(1)$, recall that the price offered during the exploitation phase of cycle c is $p^*(\widehat{\mathbf{Z}}(c))$. This price is offered to c customers, and by Corollary 2.4, the instantaneous regret incurred for each customer is $O\left(\mathbb{E}_z\left[||\mathbf{z} - \widehat{\mathbf{Z}}(c)||^2\right]\right)$. But since $\widehat{\mathbf{Z}}(c)$ is a MLE computed from c samples, it follows from a standard result that $\mathbb{E}_z\left[||\mathbf{z} - \widehat{\mathbf{Z}}(c)||^2\right] = \mathcal{O}(1/c)$. Since this prices is offered to c customers, the total regret incurred during the exploitation phase is $c \cdot \mathcal{O}(1/c) = \mathcal{O}(1)$, as claimed.

We now proceed with a rigorous proof based on the above intuition. We begin by stating a bound on the mean squared error of the maximum likelihood estimator formed by $\mathsf{MLE\text{-}CYCLE}$.

Lemma 3.7 (Mean Squared Errors for MLE based on IID Samples, Borovkov (1998))**.** *For any $c \geq 1$, let $\widehat{\mathbf{Z}}(c)$ denote the maximum likelihood estimate formed by the* $\mathsf{MLE\text{-}GREEDY}$ *policy after c exploration cycles. Then there exists a constant C_{mle} depending only on the exploration prices $\mathbf{p}$ and the problem class $\mathcal{C}$ such that*

$$\mathbb{E}_{\mathbf{z}}\left[\left\|\widehat{\mathbf{Z}}(c) - \mathbf{z}\right\|^2\right] \leq \frac{C_{mle}}{c} \ .$$

The proof of Lemma 3.7 follows from standard results on the mean-squared error of maximum-likelihood estimators, is given in detail in Appendix B. We now given the proof of Theorem 3.6.

Proof. Fix a problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ with corresponding exploration prices $\bar{\mathbf{p}}$, and consider an arbitrary cycle c. First, we show that the regret incurred during the exploration phase of cycle c is $\mathcal{O}(1)$. Since the revenue function is smooth by assumption, and since the pricing interval $\mathcal{P}$ is compact, it follows that there exists a constant $\bar{D}_1$ depending only on the problem class $\mathcal{C}$ such that

$$r(p^*(\mathbf{z}); \mathbf{z}) - r(p; \mathbf{z}) \leq \bar{D}_1$$

for all $\mathbf{z} \in \mathcal{Z}$ and all $p \in \mathcal{P}$. Consequently, the regret during the exploration phase of cycle c satisfies

$$\sum_{\ell=1}^{k} \mathbb{E}_{\mathbf{z}}[r(p^*(\mathbf{z}); \mathbf{z}) - r(\bar{p}_\ell; \mathbf{z})] \quad \leq \quad k\bar{D}_1.$$

Next, we show that the regret incurred during the exploitation phase of cycle c is also $\mathcal{O}(1)$. During the exploitation phase of cycle c, we use the greed price $p^*\left(\widehat{\mathbf{Z}}(c)\right)$, and we offer this price for c periods. It follows from Corollary 2.4 and Theorem B.1 that the instantaneous regret during the exploitation phase satisfies

$$\mathbb{E}_{\mathbf{z}}\left[r(p^*(\mathbf{z}); \mathbf{z}) - r(p^*(\widehat{\mathbf{Z}}(c)); \mathbf{z})\right] \leq c_r \; L^2 \; \mathbb{E}_{\mathbf{z}}\left[\left\|\mathbf{z} - \widehat{\mathbf{Z}}(c)\right\|^2\right] \quad \leq \quad \frac{c_r L^2 C_{mle}}{c},$$

and since the price $p^*(\widehat{\mathbf{Z}}(c))$ is offered for c periods during the exploitation phase of cycle c, we have that the total regret incurred during the exploitation phase of cycle c is bounded above by $c_r L^2 C_{mle}$. Putting everything together, we have that the cumulative regret over K cycles (corresponding to $2K + \sum_{c=1}^{K} c$ periods) satisfies

$$\text{Regret}(\mathbf{z}, \mathcal{C}, 2K + \sum_{c=1}^{K} c, \textsf{MLE-CYCLE}) \leq \left(k\bar{D}_1 + c_r L^2 C_{mle}\right) K \ .$$

Now, consider an arbitrary time period $T \geq 2$ and let $K_0 = \lceil \sqrt{2T} \rceil$. Note that the total number of time periods after K_0 cycles is at least T because $2K_0 + \sum_{c=1}^{K_0} c \geq \sum_{c=1}^{K_0} c = K_0(K_0+1)/2 \geq T$. The desired result follows from the fact that

$$\text{Regret}(\mathbf{z}, \mathcal{C}, T, \textsf{MLE-CYCLE}) \leq \text{Regret}(\mathbf{z}, \mathcal{C}, 2K_0 + \sum_{c=1}^{K_0} c, \textsf{MLE-CYCLE}).$$

□

4. The Well-Separated Case

In the general case studied in Section 3.1, there are two major obstacles to pricing that force any policy to have $\Omega(\sqrt{T})$ worst-case regret. The first obstacle is the stochastic nature of the demand. A pricing policy never observes a noise-free value of the demand curve at a given price; it observes only a random variable whose expected value is the demand at that price. The second and more prominent obstacle is that of "uninformative prices," at which no pricing policy can reduce its uncertainty about demand.

Given this observation, a natural question is the following: how much does each of the two obstacles contribute to the difficulty of dynamic pricing? More specifically, are uninformative prices so difficult to deal with that they force a minimum regret of $\Omega(\sqrt{T})$, or is it simply the stochastic nature of the demand that forces this lower bound? In this section, we shed light on this issue by considering demand curves that satisfy a "well-separated" condition (Assumption 3), which precludes the possibility of uninformative prices. Under this assumption, we show in Section 4.1 a lower bound of $\Omega(\log T)$ on the T-period cumulative regret under an arbitrary policy. Then, in Section 4.2, we show that a greedy policy achieves regret matching the order of the lower bound.

We now state Assumption 3, which guarantees that it is possible to estimate demand from customer responses at *any* price in $\mathcal{P}$.

Assumption 3 (Well Separated Assumption)**.** The problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ has a parameter set $\mathcal{Z} \subset \mathbb{R}$, and for <u>*all*</u> prices $p \in \mathcal{P}$,

(a) The family of distributions $\{Q^{p,z} : z \in \mathcal{Z}\}$ is identifiable.

(b) There exists a constant $c_f > 0$ depending only on the problem class $\mathcal{C}$ such that the *Fisher information* $I(p, z)$, given by

$$I(p,z) = \mathbb{E}_z \left[-\frac{\partial^2}{\partial z^2} \log Q^{p,z}(Y) \right]$$

satisfies $I(p, z) \geq c_f$ for all $z \in \mathcal{Z}$.

Remark 4.1 (Geometric Interpretation of Assumption 3)**.** To make the notion of well separated more concrete, one may show that any problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ satisfying Assumption 3 also has the following property: there exists some constant $c_d > 0$ depending only on $\mathcal{C}$ such that

$$|d(p; z) - d(p; \hat{z})| \geq c_d \, |z - \hat{z}|$$

for any price $p \in \mathcal{P}$, and any $z, \hat{z} \in \mathcal{Z} \subset \mathbb{R}$. We defer the details of this derivation to Appendix E.1. Thus, for any fixed price $p \in \mathcal{P}$, if we vary the demand parameter z to some other value $\hat{z}$, then the demand at price p will vary by an amount proportional to $|z - \hat{z}|$. An obvious consequence of this property and the smoothness of the demand curves is that for any two demand parameters $z \neq \hat{z}$, it must be the case that either $d(p; z) > d(p; \hat{z})$ for all $p \in \mathcal{P}$, or $d(p; z) < d(p; \hat{z})$ for all $p \in \mathcal{P}$. Thus, we refer to this condition as a "well-separated" condition, since it implies that for any two demand parameters $z \neq \hat{z}$, the corresponding demand curves do not intersect with each other.

Since we will use the maximum likelihood estimator in our pricing model and this estimator is the minimizer of the function $z \mapsto -\log\ Q_t^{\mathbf{p},z}(\mathbf{Y}_t)$, we now state Assumption 4, which gives a convenient property of the likelihood function that allows for a simple analysis of the likelihood process. As shown in Examples 4.2, 4.3, and 4.4, Assumptions 3 and 4 are satisfied by many demand families of interest, including the linear, logistic, and exponential.

Assumption 4 (Likelihood Assumptions)**.** For any sequence of prices $\mathbf{p} = (p_1, \ldots, p_t) \in \mathcal{P}^t$, the function

$$z \mapsto -\log Q_t^{\mathbf{p},z}(\mathbf{Y}_t)$$

is convex on $\mathcal{Z} \subset \mathbb{R}$.

We now state some examples of problem classes satisfying Assumptions 3 and 4.

Example 4.2 (One-Parameter Logit Family)**.** Let $\mathcal{P} = [1/2, 2]$ and let $\mathcal{Z} = [1, 2]$. Define a family of logistic demand curves by

$$d(p, z) = \frac{e^{-zp}}{1 + e^{-zp}}.$$

Then by Example 2.1, we know that this problem instances satisfies the conditions of Assumption 1. It is also straightforward to check that for any $\bar{p} \in \mathcal{P}$, the associated family $\{Q^{\bar{p},z} : z \in \mathcal{Z}\}$ is identifiable. Moreover, for any $\bar{p} \in \mathcal{P}$ and $z \in \mathcal{Z}$, we have that

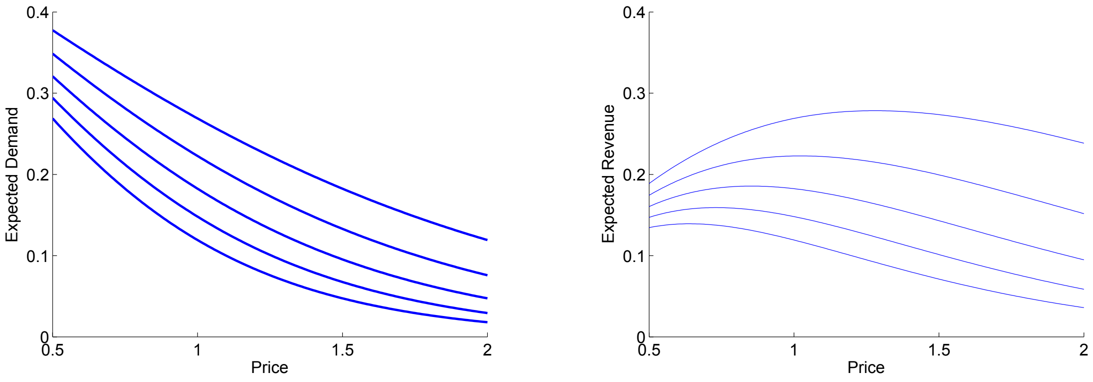
$$\frac{d}{dz} d(\bar{p}; z) = -\bar{p}\, d(\bar{p}; z)(1 - d(\bar{p}; z)) \ ,$$

and so by the formula given in Assumption 2, we have that the Fisher information is given by

$$I(\bar{p}, z) = \bar{p}^2\, d(\bar{p}; z)(1 - d(\bar{p}; z)) \geq p_{min}^2\, d_{min}(1 - d_{max}) \ .$$

Finally, it is a standard result (see, for example, Ben-Akiva and Lerman, 1985) that for the logit model, the negative log-likelihood function is globally convex, and so Assumption 4 is satisfied.

Example 4.3 (One-Parameter Linear Family)**.** Let $\mathcal{P} = [1/3, 1/2]$, let $\mathcal{Z} = [3/4, 1]$, and let $b = 2/3$ be a fixed constant. Define a linear family of demand curves by $d(p; z) = b - zp$. By Example 2.2, we know that this problem instances satisfies the conditions of Assumption 1. It is also straightforward



(a) expected demand

(b) expected revenue

Figure 2: Family of well separated logit demand and revenue curves from Example 4.2 for $z \in \{1, 5/4, 6/4, 7/4, 2\}$.

to check that for any $\bar{p} \in \mathcal{P}$, the associated family $\{Q^{\bar{p},z} : z \in \mathcal{Z}\}$ is identifiable. Moreover, for any $\bar{p} \in \mathcal{P}$ and $z \in \mathcal{Z}$, we have that

$$\frac{d}{dz} d(\bar{p}; z) = -\bar{p} \ ,$$

and we have that the Fisher information is given by

$$I(\bar{p}, z) = \frac{\bar{p}^2}{d(\bar{p}; z)(1 - d(\bar{p}; z))} \geq \frac{p_{min}^2}{d_{max}(1 - d_{min})} \ .$$

Finally, to verify Assumption 4, we have that for any vector of prices $\mathbf{p} = (p_1, \ldots, p_t) \in \mathcal{P}^t$,

$$Q_t^{\mathbf{p},\mathbf{z}}(\mathbf{y}_t) = \prod_{\ell=1}^{t} (b - zp_\ell)^{y_\ell} (1 - b + zp_\ell)^{1 - y_\ell} \ ,$$

so that

$$-\log Q_t^{\mathbf{p},\mathbf{z}}(\mathbf{y}_t) = -\sum_{\ell=1}^{t} \{y_\ell \log(b - zp_\ell) + (1 - y_\ell) \log(1 - b + zp_\ell)\} \ .$$

Taking derivatives twice, we have

$$\frac{d^2}{dz^2} \left\{-\log Q_t^{\mathbf{p},\mathbf{z}}(\mathbf{y}_t)\right\} \ = \ \sum_{\ell=1}^{t} \frac{y_\ell p_\ell^2}{(b - zp_\ell)^2} + \frac{(1 - y_\ell) p_\ell^2}{(1 - b + zp_\ell)^2} \ > \ 0 \ ,$$

which implies that the negative log-likelihood function is globally convex, as desired.

Example 4.4 (One-Parameter Exponential Family)**.** Let $\mathcal{P} = [1/2, 1]$ and let $\mathcal{Z} = [1, 2]$. Define an exponential family of demand curves by

$$d(p; z) = e^{-zp}.$$

By the same techniques used in Examples 4.2 and 4.3, one can check that this problem class satisfies all the conditions of Assumptions 1 and 3. Moreover, to verify Assumption 4, one can check that for any vector of prices $\mathbf{p} = (p_1, \ldots, p_t) \in \mathcal{P}^t$,

$$\frac{d^2}{dz^2}\left\{-\log Q_t^{\mathbf{p},\mathbf{z}}(\mathbf{y}_t)\right\} \;=\; \sum_{\ell=1}^{t} \frac{(1-y_\ell)p_\ell^2 e^{-z_\ell p}}{(1-e^{-z_\ell p})^2} \;>\; 0 \ ,$$

which implies that the negative log-likelihood function is globally convex, as desired.

4.1 A Lower Bound

In this section we establish a lower bound of $\Omega(\log T)$ for the well-separated case. The main result of this section is stated in the following theorem.

Theorem 4.5 (Well-Separated Lower Bound)**.** *Define a problem class* $\mathcal{C}_{\mathsf{WellSepLB}} = (\mathcal{P}, \mathcal{Z}, d)$ *by letting* $\mathcal{P} = [1/3, 1/2]$, $\mathcal{Z} = [2,3]$, *and letting* $d(p; z) = 1 - (pz)/2$. *Then for any policy* ψ *setting prices in* $\mathcal{P}$ *and any* $T \geq 1$, *there exists a constant* $z \in \mathcal{Z}$ *such that*

$$\mathrm{Regret}(z, \mathcal{C}_{\mathsf{WellSepLB}}, T, \psi) \geq \frac{1}{405\pi^2}\log T \ .$$

There are two key observations that lead to the proof of Theorem 4.5. First, recall that in our model, the price offered by a pricing policy ψ to the t^{th} customer is given by $P_t = \psi_t(\mathbf{Y}_{t-1})$, where $\psi_t : \{0,1\}^{t-1} \to \mathcal{P}$ is any function and $\mathbf{Y}_{t-1}$ is a vector of observed customer responses. Thus, we may think of P_t as an "estimator," since P_t is just a function ψ_t of the observed data $\mathbf{Y}_{t-1}$. Consequently, we may apply standard results on the minimum mean squared error of an estimator to show that $\mathbb{E}[(p^*(Z) - P_t)^2] = \Omega(1/t)$. We make this precise in Lemma 4.6 whose proof is given in Appendix C.

Secondly, as a converse to Corollary 2.4, we will see that it is easy to construct problem classes under which the instantaneous regret in time t is bounded *below* by the mean squared error of the price P_t with respect to the optimal price $p^*(z)$ (times some constant factors). Combining this result with the above estimate on the minimum mean squared error of P_t established the theorem.

Our proof technique follows that of Goldenshluger and Zeevi (2009), who have used van Trees' inequality to prove lower bounds on the performance of sequential decision policies.

Lemma 4.6 (Instantaneous Risk Lower Bound)**.** *Let* $\mathcal{C}_{\mathsf{WellSepLB}} = (\mathcal{P}, \mathcal{Z}, d)$ *be the problem class defined in Theorem 4.5, and let* Z *be a random variable taking values in* $\mathcal{Z}$, *with density* $\lambda : \mathcal{Z} \to \mathbb{R}_+$ *given by* $\lambda(z) = 2\{\cos(\pi(z - 5/2))\}^2$. *Then for any pricing policy* ψ *setting prices in* $\mathcal{P}$, *and for any* $t \geq 1$,

$$\mathbb{E}\left[(p^*(Z) - P_{t+1})^2\right] \geq \frac{1}{405\pi^2} \cdot \frac{1}{t},$$

where P_{t+1} is the price offered by ψ at time $t+1$, and $\mathbb{E}[\,\cdot\,]$ denotes the expectation with respect to the joint distribution of P_t and the prior density λ of the parameter $Z \in \mathcal{Z} = [2,3]$.

Here is the proof of Theorem 4.5.

Proof of Theorem 4.5. By checking first and second order optimality conditions, it is straightforward to check that $p^*(z) = 1/z$. By noting that $r'(p^*(z); z) = 0$ and $r''(p; z) = -z \leq -2$, it follows from a standard result that for any $z \in \mathcal{Z}$ and $p \in \mathcal{P}$,

$$r(p^*(z); z) - r(p; z) \geq (p^*(z) - p)^2 \ .$$

Applying this fact and Lemma 4.6, we have

$$\begin{aligned}\sup_{z \in \mathcal{Z}} \mathrm{Regret}(z, \mathcal{C}_{\mathsf{WellSepLB}}, T, \psi) \quad &\geq \quad \sup_{z \in \mathcal{Z}} \ \mathbb{E}_z \left[\sum_{t=1}^{T-1} [r(p^*(Z); Z) - r(P_{t+1}; Z)] \right] \\ &\geq \quad \mathbb{E}\left[\sum_{t=1}^{T-1} r(p^*(Z); Z) - r(P_{t+1}; Z) \right] \geq \frac{1}{405\pi^2} \sum_{t=1}^{T-1} \frac{1}{t} \ \geq \ \frac{1}{405\pi^2} \log T\end{aligned}$$

where the last line follows from the fact that $\sum_{t=1}^{T-1} \frac{1}{t} \geq \int_1^T \frac{dx}{x} = \log T$. □

4.2 A Matching Upper Bound for Well Separated Problem Class

In this section, we present a simple greedy pricing strategy called MLE-GREEDY whose regret is $\mathcal{O}(\log T)$ across all well separated problem instances, matching the order of the lower bound established in Section 4.1. We describe MLE-GREEDY in detail below, but here we sketch the intuition behind the policy.

Intuitively, we know that if we form a good estimate of the underlying demand parameter, then the optimal price corresponding to this estimate will be close to the true optimal price. More specifically, Corollary 2.4 establishes that if we compute an estimator whose mean squared error is $\mathcal{O}(1/t)$ in each time period t, then by offering the optimal prices corresponding to these estimates, we will incur instantaneous regret $\mathcal{O}(1/t)$ in each time period t, and thus incur regret that is $\mathcal{O}(\log T)$ up to time T. Thus, a natural approach is to compute an estimate of the demand parameter based on the observed customer responses to past prices offered, and then offer the best-guess optimal price corresponding to this estimate.

Although this intuition is essentially correct, there is a wrinkle to the analysis. Suppose that in time periods $1, \ldots, t$, we could observe the *actual* willingness-to-pay of each customer; that is, if we could observe the realized values $(v_1, \ldots, v_t)$ of the i.i.d. willingness-to-pay random variables $(V_1, \ldots, V_t)$. Then by standard results on maximum likelihood estimation (e.g. Theorem B.1), we could compute an estimator whose mean squared error was $\mathcal{O}(1/t)$, and by Corollary 2.4, incur regret $\mathcal{O}(1/t)$ by offering the optimal price corresponding to our estimator. However, in our model, a pricing policy does not have access to the actual willingness-to-pay of each customer. Rather, the policy observes a Bernoulli random variable $Y_t = \mathbf{1}[V_t \geq P_t]$ specifying whether the willingness-to-pay V_t of customer t exceeded the price offered P_t. Consequently, the observations $Y_1, Y_2, \ldots, Y_t$ are dependent random variables, because for any ℓ, Y_ℓ is a function of the price P_ℓ in period ℓ, which depends on the customer responses $Y_1, \ldots, Y_{\ell-1}$ in the preceding $\ell - 1$ periods. Thus, a pricing policy must form an estimate based on samples that are *dependent and not identically distributed*, and the standard bound for MLE estimates (Theorem B.1) does not apply. Thus, to establish an upper bound on the regret of MLE-GREEDY using the approach described above, it is enough to establish that the mean squared error of the estimate formed by MLE-GREEDY from t samples is in fact $\mathcal{O}(1/t)$.

With this intuition, we proceed with our analysis of the greedy pricing policy. For brevity in the following analysis, we denote by $\mathcal{G} = (\mathcal{G}_1, \mathcal{G}_2, \ldots)$ the pricing policy MLE-GREEDY described below.

Policy MLE-GREEDY$(\mathcal{C}, p_1)$

Inputs: A problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$, and an initial price $p_1 \in \mathcal{P}$.

Initialization: At time $t = 1$, offer the initial price p_1, and observe the corresponding customer decision $Y_1 = \mathbf{1}[V_1 \geq p_1]$.

Description: For time $t = 2, 3, \ldots$,

- Compute the maximum likelihood estimate $\widehat{Z}(t-1)$ given by

$$\widehat{Z}(t-1) = \arg\max_{z \in \mathcal{Z}} Q_{t-1}^{\mathcal{G},z}(\mathbf{Y}_{t-1}) \ ,$$

  where $\mathbf{Y}_{t-1} = (Y_1, \ldots, Y_{t-1})$ denotes the observed customer responses in the first $t-1$ periods.
- Offer the greedy price $p^*\left(\widehat{Z}(t-1)\right)$ based on the estimate $\widehat{Z}(t-1)$.

We now state the main result on the mean squared error on the maximum-likelihood estimator computed by MLE-GREEDY, which we prove in Appendix D.

Theorem 4.7 (MLE Deviation Inequality for Dependent Samples)**.** *Let $\widehat{Z}(t) = \arg\max_{z \in \mathcal{Z}} Q_t^{\mathcal{G},z}(\mathbf{Y}_t)$ be the maximum-likelihood estimate formed by the* MLE-GREEDY *policy. Then for any $t \geq 1$, $z \in \mathcal{Z}$, and $\epsilon \geq 0$,*

$$\Pr_z\{|\widehat{Z}(t) - z| > \epsilon\} \leq 2\, e^{-t c_H \epsilon^2/2} \qquad \textit{and} \qquad \mathbb{E}_z[(\widehat{Z}(t) - z)^2] \leq \frac{4}{c_H} \cdot \frac{1}{t}$$

The above theorem immediately yields the upper bound the regret, which is the main result of this section.

Theorem 4.8 (Well-separated Regret Upper Bound). *For any problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ satisfying Assumptions 1, 3, and 4, and any initial price $p_1 \in \mathcal{P}$, there exists a constant C_2 depending only $\mathcal{C}$ and p_1 such that for all $z \in \mathcal{Z}$ and $T \geq 2$, the* MLE-GREEDY *policy satisfies*

$$\text{Regret}(z, \mathcal{C}, T, \textsf{MLE-GREEDY}) \leq C_2 \cdot \log T \ .$$

Proof. To bound the regret incurred by MLE-GREEDY in the first period, note that since the revenue function is a smooth on the compact set $\mathcal{P} \times \mathcal{Z}$, there exists a constant $\bar{D}_2$ depending only on $\mathcal{C}$ such that $r(p^*(z); z) - r(p_1; z) \leq \bar{D}_2$ for any choice of p_1 and any $z \in \mathcal{Z}$.

To bound the regret in the subsequent periods, we apply Corollary 2.4 and Theorem 4.7 to see that

$$\begin{aligned}
\mathbb{E}_z \left[\sum_{t=1}^{T-1} r(p^*(z); z) - r(p^*(\widehat{Z}(t)); z) \right] &\leq c_r L^2 \sum_{t=1}^{T-1} \mathbb{E}_z \left[(\widehat{Z}(t) - z)^2 \right] \\
&\leq \frac{4 c_r L^2}{c_H} \sum_{t=1}^{T-1} \frac{1}{t}.
\end{aligned}$$

Taking $C_2 = \bar{D}_2 + 4 c_r L^2 c_H / (4 \log 2)$ proves the claim. □

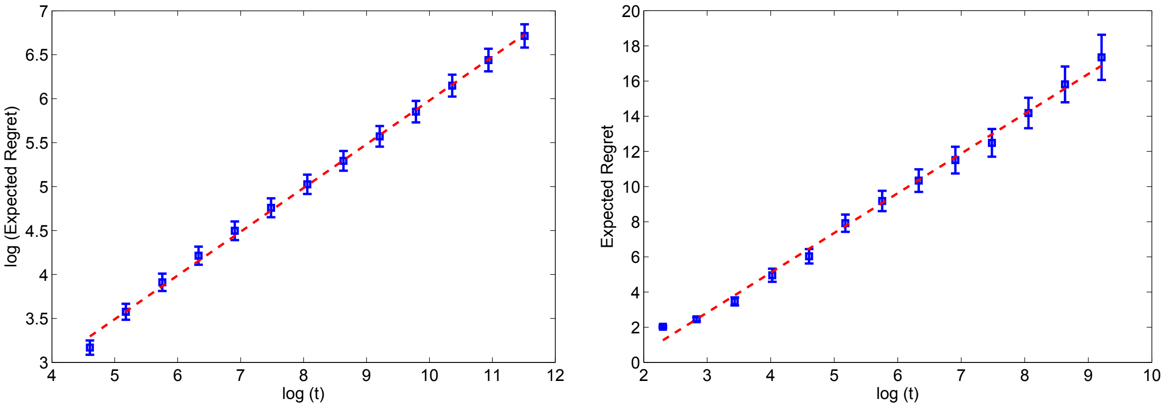
5. Numerical Experiments

In this section, we evaluate the empirical performance of the MLE-CYCLE and MLE-GREEDY policies described in Sections 3.2 and 4.2. We investigate their rates of regret, and compare their performance to the performance of several alternative policies, over a variety of problem instances. For all of our simulations, we focus on a logistic demand problem class given by $\mathcal{P} = [1/2, 8]$, $\mathcal{Z} = [0.2, 2] \times [-1, 1]$ and

$$d(p; \mathbf{z}) = \frac{e^{-z_1 p - z_2}}{1 + e^{-z_1 p - z_2}}.$$

5.1 First Simulation: Rates of Regret

For our first simulation, we investigate the rates of regret of MLE-CYCLE and MLE-GREEDY on a specific problem instance from the problem class described above. We compute the average regret of both policies over 50 independent trials for parameter values $z_1 = 1$ and $z_2 = -1$, normalizing the regret by the maximum possible per-period revenue for this instance. For MLE-CYCLE, we fix the exploration prices to be $\bar{p}_1 = 1/2$ and $\bar{p}_2 = 4.25$, corresponding to the left endpoint and midpoint of the pricing interval, and we fix the time horizon to be $T = 10^5$. For MLE-GREEDY, we fix the initial price to be $\bar{p}_1 = 4.25$, and we fix the time horizon to be $T = 10^4$.



(a) Average Regret of MLE-CYCLE

(b) Average Regret of MLE-GREEDY

Figure 3: An illustration of the rates of regret of MLE-CYCLE and MLE-GREEDY. In Figure 3 (a), the line of best fit in the log-log plot of expected regret versus T has slope 0.49, indicating that the rate of regret of MLE-CYCLE is approximately $\Theta(\sqrt{T})$. In Figure 3 (b), the expected regret of MLE-GREEDY versus $\log(T)$ is approximately linear, indicating that the rate of regret is $\Theta(\log T)$.

In Figure 3 (a), we plot the logarithm of the average regret of MLE-CYCLE versus $\log(t)$. We note that the line of best fit to the mean regret values has a slope of 0.49, which is consistent with the $\Theta(\sqrt{T})$ rate of regret established in Section 3. In Figure 3 (b), we plot the average regret of MLE-GREEDY versus $\log(t)$. The linear trend of the mean regret values is consistent with the $\Theta(\log T)$ rate of regret established in Section 4. These results provide a simple empirical example of the rates of regret of the two policies.

5.2 Second Simulation: The General Case

For our second simulation, we compare the performance of MLE-CYCLE with several alternative heuristics. We describe these alternative heuristics below.

1. FP: As a baseline for comparison, we consider a fixed-price policy FP that chooses a price uniformly at random from the pricing interval, and offers this price for all time periods. Note that this policy will have regret that is linear in T.

2. MLE-CYCLE-S: The MLE-CYCLE-S policy is a variant of MLE-CYCLE that uses samples from both the exploration *and* exploitation phases to compute its estimates of the unknown parameters (recall that the MLE-CYCLE policy computes estimates only from its explorations periods).

3. MLE-CYCLE-SU: The MLE-CYCLE-SU policy is a further refinement of MLE-CYCLE, in which all samples are used for computing the estimates, and in addition, the exploration prices are

updated at each step to be close to the estimated optimal price. Specifically, at the beginning of each cycle, we choose the first exploration price P_1 to be equal to the current estimated optimal price, and we set the second exploration price P_2 to be $P_1 + t^{-1/4}$, where t is the current time period. This scheme balances the competing objectives of having the exploration prices close to the optimal price, and having them far enough apart to provide good estimates of the demand parameters. We note that this scheme is closely related to the Controlled Variance Pricing idea introduced in den Boer and Zwart (2010).

4. KW: To compare our policies with general stochastic optimization techniques, we will consider a Kiefer-Wolfowitz-type stochastic optimization policy. Given a current price P_t, the KW policy sets

$$P_{t+1} = P_t + c_n \qquad P_{t+2} = P_t - c_n \qquad P_{t+3} = P_t + a_t \frac{Y_{t+1}P_{t+1} - Y_{t+2}P_{t+2}}{2c_t},$$

where $Y_{t+1} = \mathbf{1}[V_{t+1} \geq P_{t+1}]$ and $Y_{t+2} = \mathbf{1}[V_{t+2} \geq P_{t+2}]$. This is a stochastic gradient-ascent optimization scheme, and we implement this scheme with $a_t = t^{-1}$ and $c_t = t^{-1/4}$.

Recall that in Section 3.2, we were concerned with describing a pricing policy whose regret matched the order of the $\Omega(\sqrt{T})$ lower bound established in Section 3.1. The MLE-CYCLE policy proposed in Section 3.2 was sufficient to achieve this goal, and its simple structure facilitated a straightforward analysis of its regret, which was desirable for the theoretical development of Section 3. However, although MLE-CYCLE achieves the optimal $\mathcal{O}(\sqrt{T})$ regret, there are a number of natural modifications of this policy that one might suspect would improve its performance – specifically, the use of *all* samples to compute estimates of the demand parameters, and the updating of exploration prices as information is gained. We empirically investigate both of these modifications is this section by studying the performance of MLE-CYCLE-S and MLE-CYCLE-SU.

We investigate the performance of all pricing policies on an ensemble of problem instances drawn from a Gaussian distribution over the parameter set. We generate 500 independent random samples $(\mathbf{z}^1, \ldots, \mathbf{z}^{500})$, by drawing independent random values z_1^i in the interval $[0.2, 2]$ according to a Gaussian distribution with mean $(2 + 0.2)/2$ and variance $(2 - 0.2)/4$, truncating so that all samples lie in the interval. We then generate 500 independent random samples z_2^i for the interval $[-1, 1]$ in a similar fashion, and set $\mathbf{z}^i = (z_1^i, z_2^i)$.

To evaluate the performance of each policy, we consider the **Percentage Revenue Loss**, which is defined to be the average over the random sample of problem instances of the cumulative regret divided by the total optimal revenue. Thus, if $\mathbf{z}^1, \ldots, \mathbf{z}^m \in \mathcal{Z}$ is the sample of problem parameters, we have

$$\textbf{Percentage Revenue Loss } (T) \;\triangleq\; \frac{1}{m}\sum_{i=1}^{m} \frac{\sum_{s=1}^{T} r(p^*(\mathbf{z}^i);\mathbf{z}^i) - r(P_s^i;\mathbf{z}^i)}{T \,\cdot\, r(p^*(\mathbf{z}^i);\mathbf{z}^i)} \;\times 100\% \;.$$

Equivalently, this quantity describes the total amount of revenue lost by each policy with respect to the optimal policy, as a percentage of the total optimal revenue.

In Table 1, we report the results of these experiments. For all simulations, the exploration prices for MLE-CYCLE and its variants, and the initial price for the KW policy, are chosen uniformly at random from the pricing interval. The standard error of the figures reported in the **Percentage Revenue Loss** columns is less than 0.2% for MLE-CYCLE, MLE-CYCLE-S, and MLE-CYCLE-SU, and is less than 1.8% for FP and KW, at all reported values of T.

	Percentage Revenue Loss				
$T \times 10^3$	FP	KW	MLE-CYCLE	MLE-CYCLE-S	MLE-CYCLE-SU
1	61.9 %	58.7 %	20.4 %	14.3 %	6.0 %
2	61.9 %	58.0 %	16.1 %	10.7 %	5.0 %
3	61.9 %	57.6 %	13.9 %	9.0 %	4.5 %
4	61.9 %	57.3 %	12.5 %	7.8 %	4.2 %
5	61.9 %	57.1 %	11.5 %	7.1 %	4.0 %

Table 1: Comparison of the Percentage Revenue Loss of the heuristics on the Gaussian instance.

First, we note that all policies lose a smaller percentage of the optimal revenues than the FP policy, and more importantly, all policies have a percentage revenue loss that is decreasing with the number of time steps. We note that all three variants of MLE-CYCLE lose a significantly smaller proportion of the optimal revenue than the FP and KW policies; moreover, we see that both the use of all samples to compute the estimates of the demand parameters, as well as the updating of the exploration prices, lead to a significant improvement in the percentage revenue lost.

5.3 Third Simulation: The Well-Separated Case

As a final simulation, we will investigate the percentage revenue loss of MLE-GREEDY when problem parameters are drawn from two different distributions. Recall that to prove the lower bound of Section 4.1 on the performance of MLE-GREEDY, we showed that expected regret was $\Omega(\log T)$, when the problem parameters were drawn from a specially chosen distribution. A natural question is whether this distribution is somehow pathological, or whether the expected regret of MLE-GREEDY would be similar when problem parameters are drawn from some other type of distribution. To address this question, we generate three sets of 100 independent random problem instances for the logistic demand problem class described at the beginning of this section. Each set is generated by fixing $z_2 = 0$, and drawing z_1 from one of two distributions over the interval $[0.2, 2]$ (note that the value of $z_2 = 0$ is known to MLE-GREEDY). The first is the distribution $\frac{10}{9}\left\{\cos\left(\frac{5\pi}{9}\left(x - \frac{11}{10}\right)\right)\right\}^2$, similar to the one used in the proof of the lower bound of Section 4.1, and the second is the uniform distribution on $[0.2, 2]$. For all simulations, the starting price of MLE-GREEDY is chosen uniformly at random from the pricing interval. In Table 2, we report the percentage revenue loss of MLE-GREEDY for each of the three experiments.

	Percentage Revenue Loss			
	Lower Bound		Uniform	
$T \times 10^3$	FP	MLE-GREEDY	FP	MLE-GREEDY
1	65.2 %	1.20 %	62.3 %	1.10 %
2	65.2 %	0.67 %	62.3 %	0.61 %
3	65.2 %	0.48 %	62.3 %	0.43 %
4	65.2 %	0.37 %	62.3 %	0.34 %
5	65.2 %	0.30 %	62.3 %	0.28 %

Table 2: Comparison of the Percentage Revenue Loss of MLE-GREEDY on two distributions

The standard error for all percentage revenue loss figures reported in Table 2 is less that 0.07%. We note that the percentage revenue loss of MLE-GREEDY is much smaller than that of the fixed price policy, as well as all of the policies tested in the simulation for the general case. Moreover, we note that when averaged over 100 trials, the percentage revenue loss of MLE-GREEDY is practically identical for both problem distributions. This suggests that the lower bound distribution used in Section 4.1 is not pathological, and that we should expected similar average-case behavior for MLE-GREEDY when instances are drawn from other natural distributions.

6. Conclusion

We studied a stylized dynamic pricing problem under a general parametric choice model. For the general case, we constructed a forced-exploration policy based on maximum likelihood estimation that achieved the optimal $\mathcal{O}(\sqrt{T})$ order of regret. We also considered the special case of a "well-separated" demand family, for which a myopic maximum likelihood policy achieved the optimal $\mathcal{O}(\log T)$ order of regret. Finally, we performed an empirical investigation of the rate of regret of our policies, and compared the performance of several variations thereof. There are many possible extensions of this work, including extensions to account for the sale of multiple products and for competition among sellers. Other interesting directions would involve a more complex model of customer behavior, accounting for strategic customer decision making, or a model in which the parameter values varied over time.

References

Agrawal, R. 1995. The continuum-armed bandit problem. *SIAM Journal of Control and Optimization* **33**(6) 1926–1951.

Auer, P., N. Cesa-Bianchi, P. Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* **47**(2) 235–256.

Auer, P., R. Ortner, C. Szepesvári. 2007. Improved rates for the stochastic continuum-armed bandit problem. *20th Conference on Learning Theory (COLT)* 454–468.

Ben-Akiva, M., S. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. The MIT Press: Cambridge, MA.

Bertsimas, D., G. Perakis. 2003. Dynamic pricing: A learning approach. *Mathematical and Computational Models for Congestion Charging*, *Applied Optimization*, vol. 101. Springer, New York, 45–79.

Besbes, O., A. Zeevi. 2008. Dynamic pricing without knowing the demand function: Risk bounds and near optimal algorithms. *To appear in Operations Research* .

Besbes, O., A. Zeevi. 2009. On the minimax complexity of pricing in a changing environment. *To appear in Operations Research* .

Borovkov, A. 1998. *Mathematical Statistics*. Gordon and Breach Science Publishers.

Broadie, M., D. Cicek, A. Zeevi. 2009. General bounds and finite-time improvement for the kiefer-wolfowitz stochastic approximation algorithm. *To appear in Operations Research* .

Carvalho, A., M. Puterman. 2005. Learning and pricing in an internet environment with binomial demands. *Journal of Revenue & Pricing Management* **3**(4) 320–336.

Cope, E. 2006. Bayesian strategies for dynamic pricing in e-commerce. *Naval Research Logistics* **54**(3) 265–281.

Cope, E. 2009. Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Transactions on Automatic Control* **54** 1243–1253.

Cover, T., J. Thomas. 1999. *Elements of Information Theory*. J. Wiley, Hoboken.

den Boer, A., B. Zwart. 2010. Simultaneously learning and optimizing using controlled variance pricing. *Working paper, Centrum Wiskunde & Informatica and the University Amsterdam* .

Fabian, V. 1967. Stochastic approximation of minima with improved asymptotic speed. *Annals of Mathematical Statistics* **38**(1) 191–200.

Gallego, G., G. van Ryzin. 1994. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science* **40**(8) 999–1020.

Gill, R., B. Levit. 1995. Applications of the van trees inequality: A Bayesian Cramér-Rao bound. *Bernoulli* **1**(1) 59–79.

Goldenshluger, A., A. Zeevi. 2008. Performance limitations in bandit problems with side observations. *To appear in IEEE Transactions on Information Theory* .

Goldenshluger, A., A. Zeevi. 2009. Woodroofe's one-armed bandit problem revisited. *Annals of Applied Probability* **19**(4) 1603–1633.

Harrison, J., N. Keskin, A. Zeevi. 2010. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Working paper, Columbia and Stanford University* .

Kiefer, J., J. Wolfowitz. 1967. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics* **23**(3) 462–466.

Kleinberg, R., T. Leighton. 2003. The value of knowing a demand curve: bounds on regret for online posted-price auctions. *Proceedings of the 44th IEEE Symposium on Foundations of Computer Science*. 594–605.

Knuth, D. 1997. *The Art of Computer Programming, Volume 1: Fundamental Algorithms*. Addison-Wesley.

Lai, T., H. Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* **6**(1) 4–22.

Lim, A., J. Shanthikumar. 2007. Relative entropy, exponential utility, and robust dynamic pricing. *Operations Research* **55**(2) 198–214.

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

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

Taneja, I., P. Kumar. 2004. Relative information of type s, Csiszár's f-divergence, and information inequalities. *Information Sciences* **166**(1–4) 105–125.

Tsybakov, A. 2009. *Introduction to Nonparametric Estimation*. Springer, New York.

A. Proofs from Section 3.1

The proof of Lemmas 3.3 and 3.4 will make use of the following properties of the problem class $\mathcal{C}$ define in the statement of Theorem 3.1.

Lemma A.1 (Properties of $\mathcal{C}$)**.** *For all $p \in \mathcal{P}$ and $z \in \mathcal{Z}$,*

1. $p^*(z) = (1+2z)/(4z)$
2. $p^*(z_0) = 1$ *for* $z_0 = 1/2$.
3. $d(p^*(z_0); z) = 1/2$ *for all* $z \in \mathcal{Z}$
4. $r(p^*(z); z) - r(p; z) \geq \frac{1}{3}(p^*(z) - p)^2$
5. $|p^*(z) - p^*(z_0)| \geq \frac{1}{4}\,|z - z_0|$
6. $|d(p; z) - d(p; z_0)| \leq |p^*(z_0) - p|\ |z - z_0|$

Proof. Property 1 follows from checking first and second order optimality conditions of the revenue function $r(p; z) = pd(p; z)$. Properties 2 and 3 follow by simple calculations using the formulas for $p^*(z)$ and $d(p; z)$. Property 4 follows from the fact that $r'(p^*(z); z) = 0$ and $r''(p; z) = -2z \leq -2/3$ for all $(p, z) \in \mathcal{P} \times \mathcal{Z}$. Property 5 follows from an application of the Mean Value Theorem, and the fact that $\frac{d}{dz}p^*(z) = -1/(4z^2) \leq 1/4$ for all $z \in \mathcal{Z}$. Finally, Property 6 follows from the calculation

$$|d(p; z) - d(p; z_0)| = |1/2 + z - pz - 1/2 - z_0 + pz_0| = |z - z_0| \cdot |1 - p| = |z - z_0| \cdot |p^*(z_0) - p|$$

since $p^*(z_0) = 1$ by construction. □

The proof of Lemma 3.3 also makes use of the following standard results, which gives an upper bound on the KL-divergence between two Bernoulli distributions.

Lemma A.2 (Corollary 3.1 in Taneja and Kumar, 2004)**.** *Suppose B_1 and B_2 are distributions of Bernoulli random variables with parameters q_1 and q_2, respectively, with $q_1, q_2 \in (0,1)$. Then*

$$\mathcal{K}(B_1; B_2) \leq \frac{(q_1 - q_2)^2}{q_2(1-q_2)}.$$

A.1 Proof of Lemma 3.3

Consider a policy ψ setting prices in $\mathcal{P} = [3/4, 5/4]$ and some $s \geq 1$. To prove the lemma, we appeal to the Chain Rule for KL divergence (Theorem 2.5.3, Cover and Thomas, 1999), which states that

$$\mathcal{K}(Q_t^{\psi,z_0}; Q_t^{\psi,z}) = \sum_{s=1}^{t} \mathcal{K}(Q_s^{\psi,z_0}; Q_s^{\psi,z} | \mathbf{Y}_{s-1}),$$

where each term in the sum is the conditional KL divergence, defined as

$$\mathcal{K}(Q_s^{\psi,z_0}; Q_s^{\psi,z} | \mathbf{Y}_{s-1}) \triangleq \sum_{\mathbf{y}_s \in \{0,1\}^s} Q_s^{\psi,z_0}(\mathbf{y}_s) \log \left(\frac{Q_s^{\psi,z_0}(y_s|\mathbf{y}_{s-1})}{Q_s^{\psi,z}(y_s|\mathbf{y}_{s-1})} \right).$$

In light of this fact, we may prove the inequality of the lemma as follows. First, show that the conditional KL divergence in each time period is bounded above by the instantaneous regret in that time period (times some additional terms), and then apply the Chain Rule to show that the total KL divergence is bounded above by the cumulative regret (times additional terms).

To proceed along these lines, let $p_s = \psi(\mathbf{y}_{s-1})$. We have

$$\begin{aligned}
\mathcal{K}(Q_s^{\psi,z_0}; Q_s^{\psi,z} | \mathbf{Y}_{s-1}) &= \sum_{\mathbf{y}_s \in \{0,1\}^s} Q_s^{\psi,z_0}(\mathbf{y}_s) \log \left(\frac{Q_s^{\psi,z_0}(y_s|\mathbf{y}_{s-1})}{Q_s^{\psi,z}(y_s|\mathbf{y}_{s-1})} \right) \\
&= \sum_{\mathbf{y}_{s-1} \in \{0,1\}^{s-1}} Q_{s-1}^{\psi,z_0}(\mathbf{y}_{s-1}) \sum_{y_s \in \{0,1\}} Q_s^{\psi,z_0}(y_s|\mathbf{y}_{s-1}) \log \left(\frac{Q_s^{\psi,z_0}(y_s|\mathbf{y}_{s-1})}{Q_s^{\psi,z}(y_s|\mathbf{y}_{s-1})} \right) \\
&\leq \frac{1}{d(p_s;z)\,(1-d(p_s;z))} \sum_{\mathbf{y}_s \in \{0,1\}^{s-1}} Q_{s-1}^{\psi,z_0}(\mathbf{y}_{s-1}) \left(d(p_s;z_0) - d(p_s;z) \right)^2 , \\
&\leq \frac{3}{16} \sum_{\mathbf{y}_s \in \{0,1\}^{s-1}} Q_{s-1}^{\psi,z_0}(\mathbf{y}_{s-1}) \left(d(p_s;z_0) - d(p_s;z) \right)^2 .
\end{aligned}$$

The first line follows from the definition of conditional KL divergence. The second line follows from an algebraic manipulation using the relation $Q_s^{\psi,z_0}(\mathbf{y}_s) = Q_s^{\psi,z_0}(y_s|\mathbf{y}_{s-1}) Q_s^{\psi,z_0}(\mathbf{y}_{s-1})$, and the fact that $Q_s^{\psi,z_0}(\mathbf{y}_{s-1}) = Q_{s-1}^{\psi,z_0}(\mathbf{y}_{s-1})$. The third line follows from Lemma A.2 and the fact that

$$Q_s^{\psi,z_0}(y_s|\mathbf{y}_{s-1}) = d(p_s;z_0)^{y_s}(1-d(p_s;z_0))^{1-y_s},$$

and the fourth line follows from the fact that $d(p;z) \in [1/4, 3/4]$ for all $p \in \mathcal{P}$ and $z \in \mathcal{Z}$.

By Property 6 in Lemma A.1, we have that $(d(p_s;z_0) - d(p_s;z))^2 \leq (z_0 - z)^2 \; (p^*(z_0) - p_s)^2$, which implies

$$\begin{aligned} \mathcal{K}(Q_s^{\psi,z_0}; Q_s^{\psi,z} | \mathbf{Y}_{s-1}) &\leq \frac{3}{16} (z_0 - z)^2 \sum_{\mathbf{y}_s \in \{0,1\}^{s-1}} Q_{s-1}^{\psi,z_0}(\mathbf{y}_{s-1}) \, (p^*(z_0) - p_s)^2 \\ &= \frac{3}{16} (z_0 - z)^2 \, \mathbb{E}_{z_0} \left[(p^*(z_0) - P_s)^2 \right] \; . \end{aligned}$$

Summing over all s and using the Chain Rule for KL-divergence, we have that

$$\begin{aligned} \mathcal{K}(Q_t^{\psi,z_0}; Q_t^{\psi,z}) &= \sum_{s=1}^{t} \mathcal{K}(Q_s^{\psi,z_0}; Q_s^{\psi,z} | \mathbf{Y}_{s-1}) \;\leq\; \frac{3}{16} (z_0 - z)^2 \sum_{s=1}^{t} \mathbb{E}_{z_0} \left[(p^*(z_0) - P_s)^2 \right] \\ &\leq \frac{9}{16} (z_0 - z)^2 \sum_{s=1}^{t} \mathbb{E}_{z_0} \left[r(p^*(z_0); z_0) - r(P_s; z_0) \right] \\ &\leq \frac{9}{16} (z_0 - z)^2 \, \text{Regret}(z_0, \mathcal{C}, t, \psi) \; , \end{aligned}$$

where the last inequality follows from Property 4 in Lemma A.1. This concludes the proof.

We now proceed to the proof of Lemma 3.4. The proof of this lemma uses the following standard result on the minimal error of a two-hypothesis test, which is derived from Theorem 2.2 of Tsybakov (2009).

Lemma A.3 (Theorem 2.2, Tsybakov, 2009)**.** *Let Q_0 and Q_1 be two probability distributions on a finite space $\mathcal{Y}$, with $Q_0(y), Q_1(y) > 0$ for all $y \in \mathcal{Y}$. Then for any function $J : \mathcal{Y} \to \{0,1\}$,*

$$Q_0\{J = 1\} + Q_1\{J = 0\} \geq \frac{1}{2} e^{-\mathcal{K}(Q_0;Q_1)},$$

where $\mathcal{K}(Q_0; Q_1)$ denotes the KL divergence of Q_0 and Q_1.

A.2 Proof of Lemma 3.4

Let $z_0 = 1/2$ be as in Lemma A.1, and fix a time horizon $T \geq 2$. Let $z_1 = z_0 + \frac{1}{4} T^{-1/4}$, and define two intervals $C_{z_0} \subset \mathcal{P}$ and $C_{z_1} \subset \mathcal{P}$ by

$$C_{z_0} = \left\{ p : |p^*(z_0) - p| \leq \frac{1}{48 T^{1/4}} \right\} \;\text{ and }\; C_{z_1} = \left\{ p : |p^*(z_1) - p| \leq \frac{1}{48 T^{1/4}} \right\}.$$

Note that C_{z_0} and C_{z_1} are disjoint, since Property 5 in Lemma A.1 gives that $|p^*(z_0) - p^*(z_1)| \geq \frac{1}{4} |z_0 - z_1| = \frac{1}{16 T^{1/4}}$. It follows from Property 4 in Lemma A.1 that for each $z \in \{z_0, z_1\}$, if $p \in \mathcal{P} \backslash C_z$, then the instantaneous regret is at least $\frac{1}{3(48^2)\sqrt{T}}$ because

$$r(p^*(z); z) - r(p; z) \geq \frac{1}{3} (p - p^*(z))^2 \geq \frac{1}{3(48)^2 \sqrt{T}} = \frac{1}{3(48)^2 \cdot \sqrt{T}} \; .$$

Let $P_1, P_2, \ldots$ denote the sequence of prices under the policy ψ. Then,

$$\begin{aligned}
&\text{Regret}\,(z_0, \mathcal{C}, T, \psi) + \text{Regret}\,(z_1, \mathcal{C}, T, \psi) \\
\geq\ & \sum_{t=1}^{T-1} \mathbb{E}_{z_0}\left[r(p^*(z_0); z_0) - r(P_{t+1}; z_0)\right] + \mathbb{E}_{z_1}\left[r(p^*(z_1); z_1) - r(P_{t+1}; z_1)\right] \\
\geq\ & \frac{1}{3(48)^2 \cdot \sqrt{T}} \sum_{t=1}^{T-1} \text{Pr}_{z_0}\left\{P_{t+1} \notin C_{z_0}\right\} + \text{Pr}_{z_1}\left\{P_{t+1} \notin C_{z_1}\right\} \\
\geq\ & \frac{1}{3(48)^2 \cdot \sqrt{T}} \sum_{t=1}^{T-1} \text{Pr}_{z_0}\left\{J_{t+1} = 1\right\} + \text{Pr}_{z_1}\left\{J_{t+1} = 0\right\}\ ,
\end{aligned}$$

where for all $t \geq 1$, $J_{t+1} = \mathbf{1}[P_{t+1} \in C_{z_1}]$ is a binary random variable that takes the value of 1 when P_{t+1} is in C_{z_1}, and zero otherwise. The second inequality follows from the fact that when $J_{t+1} = 1$, we have $P_{t+1} \in C_{z_1} \subset \mathcal{P} \setminus C_{z_0}$, and thus $P_{t+1} \notin C_{z_0}$, so that $\text{Pr}_{z_0}\{J_{t+1} = 1\} \leq \text{Pr}_{z_0}\{P_{t+1} \notin C_{z_0}\}$. Now a standard result on the minimum error in a simple hypothesis test (Lemma A.3) implies that for all t,

$$\text{Pr}_{z_0}\left\{J_{t+1} = 1\right\} + \text{Pr}_{z_1}\left\{J_{t+1} = 0\right\} \geq \frac{1}{2} e^{-\mathcal{K}\left(Q_t^{\psi, z_0}; Q_t^{\psi, z_1}\right)}.$$

Now putting things together and summing over t, we have

$$\begin{aligned}
\text{Regret}(z_0, \mathcal{C}, T, \psi) + \text{Regret}(z_1, \mathcal{C}, T, \psi) &\geq \frac{1}{3(48)^2\sqrt{T}} \cdot \frac{1}{2} \sum_{t=1}^{T-1} e^{-\mathcal{K}\left(Q_t^{\psi, z_0}; Q_t^{\psi, z_1}\right)} \\
&\geq \frac{1}{3(48)^2\sqrt{T}} \cdot \frac{T-1}{2} e^{-\mathcal{K}\left(Q_T^{\psi, z_0}; Q_T^{\psi, z_1}\right)} \\
&\geq \frac{\sqrt{T}}{12(48^2)} e^{-\mathcal{K}\left(Q_T^{\psi, z_0}; Q_T^{\psi, z_1}\right)}.
\end{aligned}$$

where the second inequality follows from the standard fact that $\mathcal{K}\left(Q_t^{\psi, z_0}; Q_t^{\psi, z_1}\right)$ is non-decreasing in t (see, for example, Theorems 2.5.3 and 2.6.3 in Cover and Thomas, 1999), and the third inequality follows from the fact that $(T-1)/(2\sqrt{T}) \geq \sqrt{T}/4$ for all $T \geq 2$. This completes the proof.

B. Proof of Lemma 3.7

The proof of Lemma 3.7 is a direct application of the following standard result on the finite-sample mean-squared error of a maximum-likelihood estimator.

Theorem B.1 (Tail Inequality for MLE based on IID Samples, Theorem 36.3 in Borovkov, 1998)**.** *Let $\mathcal{Z} \subset \mathbb{R}^n$ be compact and convex, and let $\{Q^{\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ be family of distributions on a discrete sample space $\mathcal{Y}$ parameterized by $\mathcal{Z}$. Suppose Y is a random variable taking value in $\mathcal{Y}$ with distribution $Q^{\mathbf{z}}$, and the following conditions hold.*

(i) The family $\{Q^{\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ is identifiable.

(ii) For some $s > k$, $\sup_{\mathbf{z}\in\mathcal{Z}} \mathbb{E}_{\mathbf{z}}\left[\|\nabla \log Q^{\mathbf{z}}(Y)\|^s\right] = \gamma < \infty$.

(iii) The function $\mathbf{z} \mapsto \sqrt{Q^{\mathbf{z}}}$ is differentiable on $\mathcal{Z}$.

(iv) The Fisher information matrix*, whose $(i,j)^{\text{th}}$ entry is given by $\mathbb{E}_{\mathbf{z}}\left[-\frac{\partial^2}{\partial z_i \partial z_j}\log Q^{\mathbf{z}}(\mathbf{Y})\right]$, is positive definite.*

Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d. random variables taking value in $\mathcal{Y}$ with distribution $Q^{\mathbf{z}}$, and let $\widehat{\mathbf{Z}}(t) = \arg\max_{\mathbf{z}\in\mathcal{Z}} \prod_{\ell=1}^{t} Q^{\mathbf{z}}(Y_\ell)$ denote the maximum likelihood estimate based on t i.i.d. samples. Then, there exists a constants $\eta_1 > 0$ and $\eta_2 > 0$ depending only on s, k, $Q^{\mathbf{z}}$ and $\mathcal{Z}$ such that for any $t \geq 1$ and any $\epsilon \geq 0$,

$$\Pr_{\mathbf{z}}\left\{\left\|\widehat{\mathbf{Z}}(t) - \mathbf{z}\right\| \geq \epsilon\right\} \leq \eta_1\, e^{-t\eta_2\,\epsilon^2} \ .$$

To apply Theorem B.1 to our setting, we first check that the hypothesis hold for the family $\{Q^{\bar{\mathbf{p}},\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ for the exploration prices $\bar{\mathbf{p}}$ satisfying Assumption 2. For any problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$, the parameter set $\mathcal{Z}$ is compact and convex, by assumption. Conditions (i) and (iv) hold by Assumption 2, so it is enough to check conditions (ii) and (iii). To verify condition (ii), recall that for any $\mathbf{y} \in \{0,1\}^k$,

$$Q^{\bar{\mathbf{p}},\mathbf{z}}(\mathbf{y}) = \prod_{\ell=1}^{k} d(\bar{p}_\ell; \mathbf{z})^{y_\ell}(1 - d(\bar{p}_\ell; \mathbf{z}))^{1-y_\ell},$$

where $d : \mathcal{P} \times \mathcal{Z} \to [d_{min}, d_{max}]$ is smooth, with $d_{min}, d_{max} \in (0,1)$. Thus, we have

$$\nabla \log Q^{\bar{\mathbf{p}},\mathbf{z}}(\mathbf{y}) = \nabla \sum_{\ell=1}^{k} \log Q^{\bar{p}_\ell,\mathbf{z}}(y_\ell) = \sum_{\ell=1}^{k} y_\ell \nabla \log d(\bar{p}_\ell; \mathbf{z}) + (1 - y_\ell)\nabla \log(1 - d(\bar{p}_\ell; \mathbf{z})),$$

and it follows that

$$\left\|\nabla \log Q^{\bar{\mathbf{p}},\mathbf{z}}(\mathbf{y})\right\| \leq \sum_{\ell=1}^{k} \|\nabla \log d(\bar{p}_\ell; \mathbf{z})\| + \|\nabla \log(1 - d(\bar{p}_\ell; \mathbf{z}))\| \, .$$

Now since $d(\bar{p}; \cdot)$ is a smooth function that is bounded away from zero and one, we have that $\nabla \log d(\bar{p}_\ell; \mathbf{z})$ and $\nabla \log(1 - d(\bar{p}_\ell; \mathbf{z}))$ are smooth functions on the compact set $\mathcal{Z}$ for each ℓ, and it follows that there exists a constant $\bar{D}_3$ depending only on the problem instance $\mathcal{C}$ such that $\|\nabla \log Q^{\bar{\mathbf{p}},\mathbf{z}}(\mathbf{y})\| \leq \bar{D}_3$. It follows that with probability one, we have $\|\nabla \log Q^{\bar{\mathbf{p}},\mathbf{z}}(\mathbf{y})\|^s \leq \bar{D}^s$, which is the desired result.

To verify condition (iii), note that $Q^{\mathbf{p},\mathbf{z}}(\mathbf{y})$ is smooth on $\mathcal{P} \times \mathcal{Z}$, since $Q^{\mathbf{p},\mathbf{z}}(\mathbf{y})$ is a product of smooth functions on $\mathcal{P} \times \mathcal{Z}$. We also have that $Q^{\mathbf{p},\mathbf{z}}(\mathbf{y})$ is bounded away from zero, since $Q^{\mathbf{p},\mathbf{z}}(\mathbf{y}) \geq (d_{min})^k$, so it follows that $\mathbf{z} \mapsto \sqrt{Q^{\mathbf{p},\mathbf{z}}(\mathbf{y})}$ is differentiable on $\mathcal{Z}$ for any $\mathbf{p} \in \mathcal{P}^k$. Thus, we also have that $\mathbf{z} \mapsto \sqrt{Q^{\bar{\mathbf{p}},\mathbf{z}}(\mathbf{y})}$ is differentiable on $\mathcal{Z}$.

Now the result of Lemma 3.7 follows from a direct application of this theorem. Since the estimator $\widehat{\mathbf{Z}}(c)$ is formed from c i.i.d. samples, we have by Theorem B.1

$$\mathbb{E}_{\mathbf{z}}\left[\left\|\widehat{\mathbf{Z}}(c)-\mathbf{z}\right\|^2\right]=\int_0^\infty \Pr_{\mathbf{z}}\left\{\left\|\widehat{\mathbf{Z}}(c)-\mathbf{z}\right\|^2\geq u\right\}\,du\leq\int_0^\infty \eta_1 e^{-c\eta 2u}\,du=\frac{\eta_1}{c\eta_2}\ .$$

Taking $C_{mle}=\eta_1/\eta_2$ proves the claim.

C. Proof of Lemma 4.6

The proof of Lemma 4.6 depends on van Trees' inequality, which we state below.

Lemma C.1 (van Trees' Inequality, Gill and Levit, 1995)**.** *For a closed interval $\mathcal{Z}\subset\mathbb{R}$, let $\{Q^z : z\in\mathcal{Z}\}$ be a family of distributions on a discrete sample space $\mathcal{Y}$, and let Z be a random variable taking values in $\mathcal{Z}$ with density $\lambda:\mathcal{Z}\to\mathbb{R}_+$. Suppose that the following conditions hold:*

1. *For each $y\in\mathcal{Y}$, the function $z\mapsto Q^z(y)$ is absolutely continuous on $\mathcal{Z}$.*
2. *λ is absolutely continuous on $\mathcal{Z}$, and $\lambda\to 0$ at the endpoints of $\mathcal{Z}$.*
3. *$\mathbb{E}_z\left[\frac{d}{da}\log Q^z(Y)\right]=0$*

where $\mathbb{E}_z$ denotes expectation of the random variable Y having the distribution Q^z. Then, for any smooth function $g:\mathcal{Z}\to\mathbb{R}$ and any function $\hat{g}:\mathcal{Y}\to\mathbb{R}$,

$$\mathbb{E}[(\hat{g}(Y)-g(Z))^2]\geq\frac{(\mathbb{E}[\frac{d}{dz}g(Z)])^2}{\mathbb{E}\left[\left(\frac{d}{dz}\log Q^Z(Y)\right)^2\right]+\mathbb{E}\left[\left(\frac{d}{da}\log\lambda(Z)\right)^2\right]}\ , \tag{4}$$

where $\mathbb{E}[\,\cdot\,]$ denotes the expectation with respect to the joint distribution of Q^z and λ.

To apply the above result to our setting, recall the problem class $\mathcal{C}=(\mathcal{P},\mathcal{Z},d)$ defined in Theorem 4.5, which has $\mathcal{P}=[1/3,1/2]$, $\mathcal{Z}=[2,3]$, and $d(p;z)=1-(pz)/2$. For any policy ψ setting prices in $\mathcal{P}$ and any $t\geq 1$, we define the sample space to be $\mathcal{Y}=\{0,1\}^t$, and we consider the family of distributions $\left\{Q_t^{\psi,z}:z\in\mathcal{Z}\right\}$, where $Q_t^{\psi,z}:\{0,1\}^t\to[0,1]$ is the distribution of customer decisions induced by the policy ψ up to time t. That is,

$$Q_t^{\psi,z}=\prod_{\ell=1}^t(1-(p_\ell z)/2)^{y_\ell}((p_\ell z)/2)^{1-y_\ell}.$$

A convenient choice of the density $\lambda(z):[2,3]\to\mathbb{R}_+$ is $\lambda(z)=2\{\cos(\pi(z-5/2))\}^2$.

To check that the hypotheses of Lemma C.1 hold under these assumptions, note that Conditions 1 and 2 of Lemma C.1 follow immediately from our construction. Condition 3 is also satisfied because

$$\mathbb{E}_z\left[\frac{d}{dz}\log Q_t^{\psi,z}(\mathbf{Y}_t)\right]=\sum_{\mathbf{y}_t\in\{0,1\}^t}\left(\frac{\frac{d}{dz}Q_t^{\psi,z}(\mathbf{y}_t)}{Q_t^{\psi,z}(\mathbf{y}_t)}\right)Q_t^{\psi,z}(\mathbf{y}_t)=\frac{d}{dz}\sum_{\mathbf{y}_t\in\{0,1\}^t}Q_t^{\psi,z}(\mathbf{y}_t)=\frac{d}{dz}(1)=0\ .$$

By checking first and second order optimality conditions, it is straightforward to check that $p^*(z) = 1/z$, and so $p^*(z)$ is a smooth function of z on $\mathcal{Z}$. Therefore all of the conditions of Lemma C.1 are satisfied, and we can apply van Trees' Inequality to our problem.

To complete the proof, we will now compute the values on the right-hand side of van Trees' inequality (Equation 4) for our specific problem. Since $p^*(z) = 1/z$, we have that $\frac{d}{dz}p^*(z) = 1/z^2 \geq 1/9$ for all $z \in \mathcal{Z}$. It follows that $(\mathbb{E}[\frac{d}{dz}p^*(z)])^2 \geq 1/81$. Recalling that $\lambda(z) = 2\{\cos(\pi(z-5/2))\}^2$, it is straightforward to compute that

$$\mathbb{E}\left[\left(\frac{d}{dz}\log\lambda(Z)\right)\right] = 8\pi^2\int_2^3 \{\sin(\pi(z-5/2))\}^2 \, dz = 4\pi^2.$$

Finally, for any $\mathbf{z} \in \mathcal{Z}$, we may compute that

$$\mathbb{E}_z\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(\mathbf{Y}_t)\right)\middle|\mathbf{Y}_{t-1} = \mathbf{y}_{t-1}\right] = \frac{p}{z(2-pz)} \leq \frac{(1/2)}{2(2-3/2)} = \frac{1}{2},$$

where the last inequality follows from the fact that $p \in \mathcal{P}$ and $z \in \mathcal{Z}$. Applying the Chain Rule for Fisher Information (Lemma E.2), we have

$$\mathbb{E}_z\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(\mathbf{Y}_t)\right)^2\right] \leq \frac{t}{2}.$$

Since $P_{t+1} = \psi_{t+1}(\mathbf{Y}_t)$, we may apply Lemmas C.1 to get

$$\mathbb{E}[(p^*(z) - P_{t+1})^2] \geq \frac{(1/81)}{4\pi^2 + t/2} \geq \frac{1}{81(4\pi^2 + 1/2)} \cdot \frac{1}{t} \geq \frac{1}{405\pi^2} \cdot \frac{1}{t},$$

which is the desired result.

D. Proof of Theorem 4.7

In contrast to the general case, MLE-GREEDY forms an estimate of the price sensitivity parameter based on samples which are *not* i.i.d. Thus, we need to develop a new bound for our estimate. The analysis relies on the Hellinger distance. For any $t \geq 1$ and $\mathbf{y}_{t-1} \in \{0,1\}^{t-1}$, we define the conditional Hellinger distance

$$H^{\mathcal{G}}(z, u|\mathbf{y}_{t-1}) = \sum_{y_t \in \{0,1\}} \left(\sqrt{Q_t^{\mathcal{G},z}(y_t|\mathbf{y}_{t-1})} - \sqrt{Q_t^{\mathcal{G},z+u}(y_t|\mathbf{y}_{t-1})}\right)^2 ,$$

for all pairs $z \in \mathcal{Z}$ and $u \in \mathcal{Z} - z$. Note that $Q_t^{\mathcal{G},z}(y_t|\mathbf{y}_{t-1})$ denotes the probability that $Y_t = y_t$ conditioned on the event that $\mathbf{Y}_{t-1} = \mathbf{y}_{t-1}$, when the policy $\mathcal{G}$ is used and the parameter is a.

Lemma D.1 (Hellinger Distance Lower Bound)**.** *There exists a constant c_H depending only on the problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ such that for any $t \geq 1$ and any $\mathbf{y}_{t-1} \in \{0,1\}^{t-1}$, and for all pairs $z \in \mathcal{Z}$ and $u \in \mathcal{Z} - z$,*

$$H^{\mathcal{G}}(z, u|\mathbf{y}_{t-1}) \geq c_H \cdot u^2.$$

Proof. By Corollary 4.3 of Taneja and Kumar (2004), we have the following lower bound on the conditional Hellinger distance in terms of the KL divergence.

$$H^{\mathcal{G}}(z, u|\mathbf{y}_{t-1}) \geq \frac{\sqrt{d_{min}}}{2}\mathcal{K}(Q_t^{\psi,z}(\,\cdot\,|\mathbf{y}_{t-1}); Q_t^{\psi,z+u}(\,\cdot\,|\mathbf{y}_{t-1})) = \frac{\sqrt{d_{min}}}{2}\mathcal{K}(Q^{p_t,z}; Q^{p_t,z+u}),$$

where $p_t = \psi(\mathbf{y}_{t-1})$. So, to prove the desired lower bound, it is enough to prove a quadratic lower bound on the function $u \mapsto \mathcal{K}(Q^{p_t,z}; Q^{p_t,z+u})$. To do this, first note that

$$\frac{\partial^2}{\partial u^2}\mathcal{K}(Q^{p_t,z}; Q^{p_t,z+u}) = \frac{\partial^2}{\partial u^2}\mathbb{E}_z\left[\log\left(\frac{Q^{p_t,z}(Y)}{Q^{p_t,z+u}(Y)}\right)\right] = \mathbb{E}_z\left[-\frac{\partial^2}{\partial u^2}\log Q^{p_t,z+u}(Y)\right],$$

and by Assumption 3, this term is bounded below by $c_f > 0$ for all $p_t \in \mathcal{P}$ and all $z \in \mathcal{Z}$. Also, we have that

$$\frac{\partial}{\partial u}\mathcal{K}(Q^{p_t,z}; Q^{p_t,z+u})\big|_{u=0} = -\mathbb{E}_z\left[\frac{\partial}{\partial u}\log Q^{p_t,z+u}(Y)\big|_{u=0}\right] = 0$$

by a straightforward calculation. It follows from a standard result that

$$\mathcal{K}(Q^{p_t,z}; Q^{p_t,z+u}) \geq \frac{c_f}{2}u^2$$

for all $u \in \mathcal{Z} - z$. Taking $c_H = c_f\sqrt{d_{min}}/4$ proves the claim.

□

For all pairs $z \in \mathcal{Z}$ and $u \in \mathcal{Z} - z$, let the likelihood ratio $X_t^{\mathcal{G},a}(u)$ and the conditional likelihood ratio $X_t^{\mathcal{G},a}(u\big|\mathbf{Y}_{t-1})$ be defined by

$$X_t^{\mathcal{G},a}(u) = \frac{Q_t^{\mathcal{G},z+u}(\mathbf{Y}_t)}{Q_t^{\mathcal{G},z}(\mathbf{Y}_t)} \quad \text{and} \quad X_t^{\mathcal{G},a}(u\big|\mathbf{Y}_{t-1}) = \frac{Q_t^{\mathcal{G},z+u}(Y_t \mid \mathbf{Y}_{t-1})}{Q_t^{\mathcal{G},z}(Y_t \mid \mathbf{Y}_{t-1})}\,.$$

The following lemma gives an upper bound on a moment of the likelihood ratio.

Lemma D.2 (Likelihood Ratio Moment Inequality)**.** *For all pairs $z \in \mathcal{Z}$ and $u \in \mathcal{Z} - z$, and $t \geq 1$, we have*

$$\mathbb{E}_z\left[\sqrt{X_t^{\mathcal{G},z}(u|\mathbf{Y}_{t-1})}\;\middle|\;\mathbf{Y}_{t-1}\right] \leq e^{-c_H u^2/2}\,,$$

with probability one, and

$$\mathbb{E}_z\left[\sqrt{X_t^{\mathcal{G},z}(u)}\right] \leq e^{-c_H t u^2/2}.$$

Proof. To establish the first inequality, note that for all $\mathbf{y}_{t-1} \in \{0,1\}^{t-1}$,

$$\begin{aligned}
\mathbb{E}_z\left[\sqrt{X_t^{\mathcal{G},z}(u|\mathbf{Y}_{t-1})}\;\middle|\;\mathbf{Y}_{t-1} = \mathbf{y}_{t-1}\right] &= \sum_{y_t \in \{0,1\}} \sqrt{\frac{Q_t^{\mathcal{G},z+u}(y_t \mid \mathbf{y}_{t-1})}{Q_t^{\mathcal{G},z}(y_t \mid \mathbf{y}_{t-1})}} \cdot Q_t^{\mathcal{G},z}(y_t \mid \mathbf{y}_{t-1}) \\
&= \sum_{y_t \in \{0,1\}} \sqrt{Q_t^{\mathcal{G},z+u}(y_t \mid \mathbf{y}_{t-1})}\sqrt{Q_t^{\mathcal{G},z}(y_t \mid \mathbf{y}_{t-1})} \\
&= 1 - \frac{H(z,u|\mathbf{y}_{t-1})}{2} \leq e^{-H(z,u|\mathbf{y}_{t-1})/2} \leq e^{-c_H u^2/2}\,,
\end{aligned}$$

which gives the desired result. Note that the last equality follows from the definition of $H(z, u|\mathbf{y}_{t-1})$ which shows that

$$\begin{aligned} H^{\mathcal{G}}(z, u|\mathbf{y}_{t-1}) &= \sum_{y_t \in \{0,1\}} \left(\sqrt{Q_t^{\mathcal{G},a}(y_t|\mathbf{y}_{t-1})} - \sqrt{Q_t^{\mathcal{G},a+u}(y_t|\mathbf{y}_{t-1})} \right)^2 \\ &= 2\left(1 - \sum_{y_t \in \{0,1\}} \sqrt{Q_t^{\mathcal{G},a}(y_t|\mathbf{y}_{t-1})} \sqrt{Q_t^{\mathcal{G},a+u}(y_t|\mathbf{y}_{t-1})} \right) \end{aligned}$$

We will establish the second inequality of Lemma D.2 by induction on t. The case when $t = 1$ follows immediately from the above calculation. So, assume the claim holds for $t - 1$, that is,

$$\mathbb{E}_z \left[\sqrt{X_{t-1}^{\mathcal{G},z}(u)} \right] \leq e^{-(t-1)c_H u^2/2}$$

Now, by definition, we have that

$$\begin{aligned} \mathbb{E}_z \left[\sqrt{X_t^{\mathcal{G},z}(u)} \right] &= \mathbb{E}_z \left[\sqrt{X_{t-1}^{\mathcal{G},z}(u)} \cdot \sqrt{X_t^{\mathcal{G},z}(u|\mathbf{Y}_{t-1})} \right] \\ &= \mathbb{E}_z \left[\sqrt{X_{t-1}^{\mathcal{G},z}(u)} \cdot \mathbb{E}_z \left[\sqrt{X_t^{\mathcal{G},z}(u|\mathbf{Y}_{t-1})} \;\middle|\; \mathbf{Y}_{t-1} \right] \right] \\ &\leq e^{-c_H u^2/2} \cdot \mathbb{E}_z \left[\sqrt{\mathbb{E}_z[X_{t-1}^{\mathcal{G},z}(u)]} \right] \leq e^{-tc_H u^2/2} , \end{aligned}$$

where the first inequality follows from the first part of Lemma D.2, and the final inequality follows from the inductive hypothesis. This completes the proof. □

Here is the proof of Theorem 4.7.

Proof. Consider an arbitrary $a \in \mathcal{Z}$. For all $u \in \mathcal{Z} - z$, let $L_t^{\mathcal{G},z}(u) = -\log X_t^{\mathcal{G},z}(u)$. By Assumption 4, $L_t^{\mathcal{G},z}(u)$ is globally convex in u. Moreover, it is easy to verify that $L_t^{\mathcal{G},z}(0) = 0$. It follows from the definition of $\widehat{Z}(t)$ that

$$\widehat{Z}(t) = \arg\max_{v \in \mathcal{Z}} Q_t^{\mathcal{G},v}(\mathbf{Y}_t) = z + \arg\max_{u \in \mathcal{Z}-z} X_t^{\mathcal{G},a}(u) = z + \arg\min_{u \in \mathcal{Z}-z} L_t^{\mathcal{G},a}(u)$$

Therefore, for any $\delta \in \mathcal{Z} - z$, if $|\widehat{Z}(t) - z| > |\delta|$, then the minimizer of $L_t^{\mathcal{G},z}(\,\cdot\,)$ must be outside the interval $[-\delta, \delta]$, which implies that either $L_t^{\mathcal{G},z}(\delta) \leq 0$ or $L_t^{\mathcal{G},z}(-\delta) \leq 0$. Hence, for any $\delta \in \mathcal{Z} - z$, we have that

$$\Pr_z\{|\widehat{Z}(t) - z| \geq |\delta|\} \leq \Pr_z\{L_t^{\mathcal{G},z}(\delta) \leq 0\} + \Pr_z\{L_t^{\mathcal{G},z}(-\delta) \leq 0\} .$$

By Markov's Inequality and Lemma D.2, it follows that

$$\begin{aligned} \Pr_z\{L_t^{\mathcal{G},z}(\delta) \leq 0\} &= \Pr_z\{X_t^{\mathcal{G},z}(\delta) \geq 1\} = \Pr_z\left\{ \sqrt{X_t^{\mathcal{G},z}(\delta)} \;\geq\; 1 \right\} \\ &\leq \mathbb{E}_z \left[\sqrt{X_t^{\mathcal{G},z}(\delta)} \right] \;\leq\; e^{-tc_H \delta^2/2} \quad . \end{aligned}$$

A similar argument shows that $\Pr_z\{L_t^{\mathcal{G},z}(-\delta) < 0\} \leq e^{-tc_H\delta^2/2}$, which implies that for any $\delta \in \mathcal{Z} - z$,

$$\Pr_z\{|\widehat{Z}(t) - z| > |\delta|\} \leq 2\, e^{-tc_H\delta^2/2} \ .$$

Thus, for any $0 < \epsilon \leq \max\{|x| : x \in \mathcal{Z} - z\}$, we have that

$$\Pr_z\{|\widehat{Z}(t) - z| > \epsilon\} \leq 2e^{-tc_H\epsilon^2/2} \ .$$

On the other hand, if $\epsilon > \max\{|x| : x \in \mathcal{Z} - z\}$, then $\Pr_z\{|\widehat{Z}(t) - z| > \epsilon\} = 0$ by definition. This gives the desired result.

The upper bound on the mean squared error follows immediately because

$$\mathbb{E}_z[(\widehat{Z}(t) - z)^2] = \int_0^\infty \Pr_z\{(\widehat{Z}(t) - z)^2 > u\}\ du \ \leq \ 2\int_0^\infty e^{-tc_H\ u/2}\ du \ = \ \frac{4}{c_H} \cdot \frac{1}{t}$$

□

E. Proofs of Auxiliary Results

E.1 Proof of Remark 4.1

By Lemma A.2, we have that

$$(d(p; z) - d(p; z + u))^2 \geq d_{min}(1 - d_{max})\mathcal{K}(Q^{p,z}; Q^{p,z+u}) \ .$$

Now by applying the arguments of Lemma D.1 and using Assumption 3, we have that

$$\mathcal{K}(Q^{p,z}; Q^{p,z+u}) \geq \frac{c_f}{2}u^2.$$

Choosing $c_d = d_{min}(1 - d_{max})c_f/2$ establishes the inequality.

E.2 Chain Rule for Fisher Information

It is a standard result (e.g. Cover and Thomas, 1999, Exercise 11.19) that for distributions satisfying mild regularity assumptions (which are satisfied in our model), the Fisher information may also be written as

$$\mathbb{E}_z\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(\mathbf{Y}_t)\right)^2\right] = -\mathbb{E}_z\left[\frac{d^2}{dz^2}\log Q_t^{\psi,z}(\mathbf{Y}_t)\right],$$

So it follows that

$$\begin{aligned}
\mathbb{E}_z\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(\mathbf{Y}_t)\right)^2\right] &= -\mathbb{E}_z\left[\frac{d^2}{dz^2}\log\prod_{\ell=1}^{t} Q_t^{\psi,z}(Y_\ell \mid \mathbf{Y}_{\ell-1})\right] \\
&= \sum_{\ell=1}^{t} -\mathbb{E}_z\left[\frac{d^2}{dz^2}\log Q_t^{\psi,z}(Y_\ell \mid \mathbf{Y}_{\ell-1})\right] = \sum_{\ell=1}^{t}\mathbb{E}_z\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(Y_\ell \mid \mathbf{Y}_{\ell-1})\right)^2\right] .
\end{aligned}$$

This completes the proof.