Dynamic Pricing and Demand Learning with Limited Price Experimentation

David Simchi-Levi* He Wang† Alexander M. Weinstein‡

May 1, 2014

Abstract

In a dynamic pricing problem where the demand function is not known a priori, price experimentation can be used as a demand learning tool. Existing literature often assumes no constraint on price changes, but in practice sellers are faced with business constraints that prevent them from conducting extensive experimentation. We consider a dynamic pricing model where the demand function is unknown but has finite possibilities. The seller can change price to learn demand, but faces a constraint on the number of price changes during the sales window. We characterize the impact of the price change constraint on the seller's revenue. Specifically, if the seller changes price only once, the regret compared to the full information case is of order $\Theta(\log n)$. By changing price m times, the regret can be further reduced to $\Theta(\log^{(m)} n)$, or m iterations of the logarithm. This characterization provides new insight into the result by Harrison et al. (2012), which shows the regret is bounded by a constant if there is no constraint on price changes.

1 Introduction

The increased availability of customer data in recent years has made possible new approaches to pricing for revenue maximization in settings of incomplete information about demand. These learning-and-optimizing approaches to dynamic pricing can be used when the underlying relationship between price and demand is fixed over time but the exact functional relationship is

*MIT, email: dslevi@mit.edu

†MIT, email: wanghe@mit.edu

‡MIT, email: amw22@mit.edu

initially unknown. Sellers can extract information from the customer data "on the fly" and use this information to price for optimal revenue.

Online retailing is an important example of this data revolution, as sellers are now able to capture data in real time when customers log on and make purchases. Furthermore, by selling products on the web, sellers can change price at essentially no cost, which provides greater opportunities for price experimentation. At least in theory, dynamic pricing with price experimentation has proven to increase expected revenue. In fact, most papers we list in Section 2 show that, under various model assumptions, by incorporating demand learning into dynamic pricing policies, sellers can achieve revenue very close to that of the full information case.

However, this learning-and-optimizing approach is not commonly used in practice. For example, Harrison et al. (2012) observe that price experimentation is rare in financial services, even if there is obvious opportunity for customized pricing. We have had the same observation through collaboration with two e-commerce companies[1]—companies that collect customer purchasing information in real time but do not use them to set prices. This observation shows that, despite the reduction of physical cost of changing price, business constraints still makes it difficult to implement price experimentation.

The first business constraint is a short sales windows. For example, fashion apparel products typically have short life cycles. As a result, while simple mark-up or mark-down policies are possible, it is difficult to carry out an extensive price experiment for fashion products within the sales horizon (see Feng and Gallego 1995, Caro and Gallien 2012). Another example is in "flash sale" websites, which offer limited-time discounts on various products, for which the sales window is often measured in days, if not hours. The extremely short sales window, plus web retailers' policy that customers must be offered a single price at any time, makes it very difficult to conduct price experimentation. But there is still great incentive to do it: Data show that flash sale sites have grown rapidly, with reported annual revenue growth of 50% over the past five years (Wolverson 2012).

The second business constraint is the reaction from customers. Excessive price changes may cause confusion and affect the seller's brand reputation. Dynamic pricing also leads to strategic customer behaviors, which may either increase or decrease the seller's revenue. A discussion of how dynamic pricing policies affect customers' strategic behavior is beyond the scope of this paper; readers are referred to the survey by Aviv et al. (2009). However, the intuition of our two e-commerce partners is that fewer price changes cause less confusion and less strategic customer

[1] Due to privacy issues, we cannot reveal the identity of the two companies.

behavior. Therefore, it is to sellers' advantage to limit the number of prices offered, provided that sufficient demand learning can be attained to maximize revenues.

Motivated by these practical constraints on price experimentation, we consider a dynamic pricing model where there is an explicit constraint on the number of price changes. In a Bayesian setting, we assume the true demand function belongs to a finite set of possible demand functions, and the seller updates the posterior probability by observing customer purchase history. This setting is closely related to the model considered by Harrison et al. (2012), where they assume two possible demand functions and no constraint on price changes. However, we will show that the asymptotic behavior of the model is dramatically different when we add the constraint requiring the number of price changes bounded by a finite number.

In our main result, we quantify the impact on revenue if such a price change constraint is present. The seller's revenue performance is measured by *regret*, defined as the gap between the revenue achieved by limited price changes and the revenue of a clairvoyant who has full information on demand function. When there are n customer arrivals during the sales horizon, we show the regret is of order $O(\log n)$ if the seller changes price only once during the sales window. By changing price m times, the regret can be further reduced to $O(\log^{(m)} n)$, or m iterations of the logarithm. These regret bounds are tight, because they match the $\Omega(\log^{(m)} n)$ lower bound on regret of *any* pricing policy, proved in the companion work by Cheung and Simchi-Levi (2014).

Interestingly, Harrison et al. (2012) show that the regret can be bounded by a constant using a myopic pricing policy. The myopic policy resets the price after each period to maximize revenue for the next period. Comparing Harrison et al.'s constant regret bound to the iterated logarithm bound that we derived, we immediately have the following observation: First, imposing the price change limitation always incurs a cost since the seller cannot achieve the constant regret asymptotically by using any *finite* number of price changes. Second, the incremental effect of price changes decreases extremely fast. The first price change reduces regret from $\Theta(n)$ to $\Theta(\log n)$; each additional price change thereafter compounds a logarithm to the order of regret. As a result, the first few price changes generate most of the benefit of dynamic pricing. This property allows sellers to design pricing policies that satisfy the business constraints on price experimentation mentioned previously, while still achieving good revenue performance.

The remainder of this paper is organized as follows. §2 reviews related literature. §3 defines the mathematical model of the problem considered. The main result on asymptotic regret bounds is presented in §4. In §5, we consider an extension where inventory is limited and the

number of customer arrivals is random. The regret is bounded by $O(\sqrt{n})$ for any number of price changes, where n represents both the scale of customers and the scale of inventory. Numerical examples are shown in §6.

2 Related Literature

Dynamic pricing with demand learning has been a popular research area over the last decade. The research interest is driven by industry practice where e-commerce makes it possible to analyze sales data and adjust price in real time. The problem is also interesting in theory because of the inherent exploration-exploitation trade-off. For a comprehensive survey on learning-and-optimizing approaches to dynamic pricing, readers are referred to the book chapter by Aviv and Vulcano (2012) and a recent paper by den Boer (2014a).

To address the exploration-exploitation trade-off, one approach is to fit the problem to the well-studied multi-armed bandits setting. Kleinberg and Leighton (2003) show that if the interval of allowable prices is carefully discretized into finite arms, then algorithms for the classical multi-armed bandits problems such as UCB and Exp3 can be directly applied. Mersereau et al. (2009) and Rusmevichientong and Tsitsiklis (2010) consider a bandits problem where the rewards of different arms are correlated by a linear function. Their models can be directly applied to a pricing setting in which demand is linear but the functional parameters are unknown. Badanidiyuru et al. (2013) consider a multi-armed bandits model with resource constraints.

When inventory is limited, a number of papers apply settings that follow the classical model by Gallego and Van Ryzin (1994), which considers a vendor selling a given stock of inventory over a finite time period. Aviv and Pazgal (2005) were among the first to consider incomplete information in this model setting, where they assume the seller has a prior belief on the arrival rate and keeps updating his belief based on sales data. Their work is extended by Araman and Caldentey (2009) and Farias and Van Roy (2010). Both papers consider infinite time-horizon problems and different assumptions on the prior distribution of arrival rate. Besbes and Zeevi (2009) consider a finite horizon model and allow the reservation price distribution (demand function) to be unknown as well. They focus on two cases: the parametric case where the functional form of the demand function is given but the parameter values are unknown, and the nonparametric case where the demand function is only assumed to be Lipschitz continuous. A similar nonparametric model is considered by Wang et al. (2014). By further assuming that the demand function is twice differentiable, they give an efficient algorithm that mixes the learning and earning phases.

Unlike the above literature, we assume that the seller can observe all customers' purchase decisions, not only those that result in sales. The same assumption is made by Carvalho and Puterman (2005) and Cope (2007). Both papers consider e-commerce settings using discrete-time models. They assume that customer arrivals per period are time-homogenous and independent of price, but the probability that a customer makes a purchase (also known as the demand function) depends on price and this relationship is unknown to the seller. Carvalho and Puterman (2005) assume that arrivals per period follow a Poisson distribution, and the demand function belongs to the logit family. Cope (2007) assumes a Dirichlet distribution for the demand function but restricts allowable prices to a finite set.

Recently, a stream of papers has considered several variations of myopic (or greedy) and semi-myopic pricing policies. Broder and Rusmevichientong (2012) considers a seller who repeatedly updates demand information by maximum-likelihood estimation, while den Boer and Zwart (2014), den Boer (2014b) applies what they called controlled variance pricing using maximum quasi-likelihood estimation. Harrison et al. (2012) uses a myopic Bayesian updating process in which two demand hypotheses, i.e., two demand functions, are tested. Keskin and Zeevi (2013) derives a greedy policy based on least-squares estimation. Besbes and Zeevi (2013) show that nonlinear demand can be effectively approximated by linear functions using myopic policy. The idea shared by these papers is that when demand information is updated at each time period and price is chosen myopically to maximize revenue, learning will take care of itself, and thus no active price exploration is needed. In fact, all of these papers show that myopic or near-myopic algorithms have provable regret bounds.

Our model is closely related to the one considered in Harrison et al. (2012). Although Harrison et al. (2012) assume two demand hypotheses while we allow for any finite hypotheses, we believe this difference is nonessential in theory and it is possible to extend their analysis to the finite hypotheses case. However, a major difference is that the model in Harrison et al. (2012) does not limit the number of price changes, which motivates them to consider a myopic revenue-maximizing policy in which each customer is offered a distinct price. In our model, because of the existence of price changing constraint, we need to consider a different family of policies. We also show that the constant regret bound in Harrison et al. (2012) is the limitbut cannot be achieved by any finite number of price changes.

In dynamic pricing literature, several papers consider explicit limitation on price changes in settings assuming complete information, e.g. Feng and Gallego (1995), Bitran and Mondschein (1997), Netessine (2006). For management practice, Caro and Gallien (2012) report that fashion

company Zara uses a clearance pricing policy with a pre-determined price set, which inherently allows for only a limited number of price mark-downs. Zbaracki et al. (2004) provide empirical evidence on the cost of price changing. In the learning-and-pricing literature, the only paper that we are aware of considering price changing constraint is the thesis by Broder (2011). The thesis finds the minimal number of prices needed to achieve a regret bound that is in the same order of the regret bound under unconstrained price changes. By contrast, we take a direct approach by explicitly including the number of price changes as a hard constraint, and characterize how the order of the regret bound decreases with the number of price changes.

3 Problem Formulation and Assumptions

We consider a monopolist offering a single product during a web-hosted sales event with finite time horizon. When the event starts, there are s units of inventory, which may not be replenished during the event. In a sequential fashion, customers log on to the website and decide whether or not to purchase the item.

The sales horizon is divided into n periods so that each time period has at most one customer arrival. We define random variables

$$Y_t := \begin{cases} 1, & \text{if there is a customer logging on at period } t, \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

for period $t = 1, \cdots, n$. We assume that Y_t are independent and identically distributed with mean y. [2] Y_t is also independent of price, because we assume that customers see the price after logging on to the website. Based on historical data from similar events, the seller knows *a priori* the distribution of Y_t.

The seller chooses price p from a given set $\mathcal{P}$. Because the product being sold has never been offered before, the seller does not know the demand. However, based on previous sales of similar products, the latent demand function is known to be a component of the vector function $\varphi := (\varphi_1(p), \ldots, \varphi_k(p))^T$, where each hypothetical demand relationship $\varphi_i : \mathcal{P} \to [0, 1]$ is defined by the *probability* that a customer purchases the item at a given unit price p. A similar model with $k = 2$ is considered by Harrison et al. (2012). We say hypothesis i holds if the actual demand function is $\tilde{\varphi}(p) = \varphi_i(p)$. The actual demand function does not change over time.

As customers log on, the seller can observe in real time whether the customer decides to

[2] If the arrival rate, namely the mean of Y_t, is time varying, we can take the length of each time period reciprocal to the arrival rate, and reduce the problem to the i.i.d. case.

purchase the item or log off without making a purchase. The customer purchase decisions are modeled as independent Bernoulli variables X_t with mean $\tilde{\varphi}(p)$, where

$$X_t := \begin{cases} 1, & \text{if the customer at period t purchases the item} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

We assume that each customer reacts only to the current price offered. [3]

At the end of each time period, the seller decides whether to maintain the current price or change the price. Due to business constraints, the seller can make only a limited number of changes to the price over the course of the sales event. Unless specified otherwise, our analysis is focused on the basic case where only one price change is allowed. Pricing policies satisfying this latter constraint are referred to as *two-price policies*. Similarly, policies that use three distinct prices are referred to as three-price policies, etc.

The probability vector $q := (q^1, \ldots, q^k)^T$ represents the seller's belief that each corresponding component of φ is the actual latent demand, where $q \in Q := \{(q^1, \ldots, q^k) | 0 \leq q^i \leq 1, \sum_{i=1}^k q^i = 1\}$. For any given belief probability vector q, we define the expected demand $\sigma(p, q) := \varphi^T q$, where φ^T represents the transpose of the column vector φ. Note that $\sigma(p, q)$ is linear in q and also depends on price p. For a single customer, the expected revenue is $r(p, q) := p\sigma(p, q)$. The maximum expected revenue is $r^*(q) = \max_{p \in \mathcal{P}} r(p, q)$. We also assume the maximizing price $p^*(q) \in \arg\max_{p \in \mathcal{P}} r(p, q)$ exists for all q.

Based on expertise and historical data from similar products, the seller has some prior belief vector, $q = q_0$. If a customer shows up at period t ($Y_t = 1$), after his purchase decision X_t is realized under price p, the seller updates posterior belief probabilities using Bayes's rule:

$$q_t^i = \begin{cases} \alpha_i(q_{t-1}) := q_{t-1}^i \dfrac{\varphi_i(p)}{\sigma(p, q_{t-1})}, & \text{if } Y_t = 1, X_t = 1, \\ \beta_i(q_{t-1}) := q_{t-1}^i \dfrac{1 - \varphi_i(p)}{1 - \sigma(p, q_{t-1})}, & \text{if } Y_t = 1, X_t = 0, \\ q_{t-1}^i, & \text{if } Y_t = 0 \end{cases} \tag{3}$$

where q_t^i is the i-th component of q_t. We also define vector functions $\alpha(q) = (\alpha_1(q), \ldots, \alpha_k(q))^T$ and $\beta(q) = (\beta_1(q), \ldots, \beta_k(q))^T$.

[3] Because the seller is free to increase or decrease price, there is no incentive for strategic customer waiting. For a detailed review of existing literature desribing demand learning in the presence of strategic consumer behavior see Aviv and Vulcano (2012). While the learning algorithm could be susceptible to misleading consumer behavior during the learning phase, we assume security measures can counteract the effect of such behavior. For example, the system could identify and block users who log in repeatedly without purchasing the item in an attempt to bluff the algorithm into lowering the price.

3.1 Bayesian Regret of Policies

A Bayesian dynamic pricing policy $\pi = (\pi_1, \ldots, \pi_n)$ sets the price at period t according to $p_t = \pi_t(q_{t-1}, s_{t-1})$, where q_{t-1} and s_{t-1} are the belief probability and inventory level at the end of previous period.[4] If hypothesis i holds, the probability measure induced by policy π is given by

$$\mathbb{P}_i^\pi(Y_1 = y_1, X_1 = x_1, \cdots, Y_n = y_n, X_n = x_n) = \prod_{t=1}^{n} y^{y_t}[1-y]^{1-y_t}[\varphi_i(p_t)]^{x_t}[1-\varphi_i(p_t)]^{1-x_t}, \quad (4)$$

where $p_1, \ldots, p_n$ is the price sequence determined by π for sales realizations $x_1, \ldots, x_n$, and y is the mean of Y_t. [5]

We denote by $\mathbb{E}_i^\pi(\cdot)$ the expectation operator associated with the probability measure $\mathbb{P}_i^\pi(\cdot)$. Under a given policy π and inventory level s, the expected revenue conditional on hypothesis i is given by

$$R_{i,n}^\pi(s) = \mathbb{E}_i^\pi\left[\sum_{t=1}^{n} p_t X_t Y_t\right]. \quad (5)$$

The expected revenue given a belief probability q is

$$R_n^\pi(q, s) = \sum_{i=1}^{k} q^i R_{i,n}^\pi(s),$$

and the seller's objective is to maximize the total expected revenue $R_n^*(q, s) = \max_\pi R_n^\pi(q, s)$ given that $q = q_0$. The maximization is taken over all policies that change price at most once.

It is useful to interpret the objective function from another perspective. We define

$$\Delta_{i,n}^\pi(s) = \tilde{R}_{i,n}(s) - R_{i,n}^\pi(s) \quad (6)$$

as the *regret* for hypothesis i, where $\tilde{R}_{i,n}(s)$ the the expected revenue of a clairvoyant who knows the correct demand function $\tilde{\varphi}$. Therefore, without knowing the latent demand function, the seller can never do better than the clairvoyant. Thus, the quantity in Equation (6) is a positive quantity, representing the gap between what the clairvoyant would achieve and what the seller

[4] When customer arrival rate is high, it is possible that a customer arrives before the previous customer makes the purchase decision. In that case, p_t cannot depend on q_{t-1}, but on some $q_{t-\tau}$, where $\tau > 1$ is the delay time. However, such delay would not have a big effect when $\tau \ll n$.

[5] Strictly speaking, X_t has meaning only when $Y_t = 1$, but defining it for all periods does not change the objective function (5) and simplifies the notation.

can achieve by using policy π. The expected regret under policy π is defined as

$$\Delta_n^\pi(q,s) = \sum_{i=1}^{k} q^i \Delta_{i,n}^\pi(s),$$

and the minimum expected regret is $\Delta_n^*(q,s) = \min_\pi \Delta_n^\pi(q,s)$. Because the clairvoyant's total expected revenue is fixed, minimizing expected regret is equivalent to maximizing expected revenue.

4 Basic Model: Deterministic Customer Arrival, Unlimited Inventory

We present our main result in this section. We start by giving a two-price policy that achieves an upper bound on regret of order $O(\log n)$, and use it to construct a policy that changes price m times and achieves an upper bound of $O(\log^{(m)} n)$. In the companion work of this paper, Cheung and Simchi-Levi (2014) show the matching lower bounds of $\Omega(\log n)$ and $\Omega(\log^{(m)} n)$ on any policies with one and m price changes, respectively. Thus, our polices achieve the optimal regret bounds up to a constant.

We consider a basic version of the model. First, we assume that the seller knows the exact number of customers logging on during the sales event. Therefore, each customer arrival can be modeled as a distinct period. In other words, we let $Y_t = 1$ for all $t = 1, \cdots, n$ in the definition (1). Second, we assume that inventory is sufficient (e.g. $s \geq n$). As a result, the seller's pricing policy $\pi_t(q_{t-1})$ only depends on the belief probability q_{t-1}, but not on the inventory level s_{t-1}. In addition, we drop the s variable of revenue and regret functions. We will see later in section 5 that the first assumption is not essential. If the customer arrival is random and inventory is sufficient (to be defined in section 5), the seller can still achieve a regret bound of $O(\log n)$ with a two-price policy. It is simple to extend the analysis for a multiple price policy, constructed in the same manner as in this section, that achieves the $O(\log^{(m)} n)$ bound. However, the second assumption is essential. Our analysis shows that limited inventory leads to a regret of $O(\sqrt{n})$, which dominates the price changing effect, so the characterization cannot be derived in a limited inventory setting.

4.1 Effectiveness of Limited Price Experimentation

When inventory is unconstrained, the regret function defined in equation (6) can be rewritten as

$$\Delta_{i,n}^{\pi} = nr^*(e_i) - R_{i,n}^{\pi} \tag{7}$$

where e_i is a unit vector with all but the i-th components equal to zero. To understand the meaning of Equation (7), note that if inventory is sufficient, a clairvoyant who knows φ_i is the actual demand function will choose the revenue maximizing price $p^*(e_i)$ for every period. Therefore, the clairvoyant is able to achieve the maximum expected revenue $r^*(e_i)$ for each period, so the first term is the clairvoyant's total expected revenue under hypothesis i.

This section presents our main result: a two-price policy, and in general policies with few price changes, has good performance in maximizing revenue. Throughout this section, we assume that the initial belief probability satisfies $q_0^i > 0, \forall i = 1, \ldots, k$. If there exists $q_0^i = 0$, we know hypothesis i will never happen and we can simply exclude demand function $\varphi_i(p)$ from the model. We say a price p is *discriminative* if $\varphi_i(p) \neq \varphi_j(p)$ for any two different demand hypotheses i and j. Intuitively, if a price is not discriminative, one cannot identify the latent demand function under this price. The following lemma shows the converse: if a price is discriminative, then one can identify the latent demand function correctly and efficiently.

Lemma 1. *Suppose policy π keeps a fixed discriminative price p until period t. Then there exists a constant $\lambda > 0$ such that*

$$\mathbb{E}_i^{\pi}[q_t^i] \geq 1 - (2k/q_0^i)\exp(-\lambda t) \tag{8}$$

for any hypothesis $i = 1, \ldots, k$. The constant λ depends only on the demand function φ and price p.

The lemma implies that a fixed price policy is effective for demand learning in the following sense: If hypothesis i holds, the seller's belief that hypothesis i is true will converge to one as he keeps learning from customer purchase decisions, and the convergence rate is exponential.

To show a regret bound for two-price policies, consider the following heuristic pricing policy.

Heuristic h: Choose a discriminative initial price p and apply it for $T = \lfloor \frac{1}{\lambda} \log n \rfloor + 1$ periods, where constant λ is given by Lemma 1. If the i-th component of belief probability q_T satisfies $q_T^i \geq 1/k$, apply price $p^*(e_i)$ for the remaining periods. (Note that $\sum_{i=1}^k q_T^i = 1$, so such component always exists.) If there are multiple components satisfying this condition,

choose randomly among them.

The heuristic divides the sales horizon into two parts: a learning phase with a fixed number of periods, where an initial price is tested and demand information is collected; then an optimization phase where the most probable demand function is used. We are now ready to establish the main result of this section.

Theorem 1. *Suppose there exists a discriminative price in price set $\mathcal{P}$, then for $n \geq 2$ there exists a constant C such that for any belief probability $q_0 \in Q$, $\Delta_n^*(q_0) \leq \Delta_n^h(q_0) \leq C \log n$.*

Proof. By definition, $\Delta_n^*(q_0) \leq \Delta_n^h(q_0)$ because the former is the regret of the optimal policy, so we only need to show the second inequality.

If hypothesis i is true, consider the regret for hypothesis i:

$$\Delta_{i,n}^h = \Delta_{i,\text{learn}}^h + \Delta_{i,\text{optimize}}^h, \tag{9}$$

where $\Delta_{i,\text{learn}}^h$ is the total expected regret from period 1 to T (learning phase), $\Delta_{i,\text{optimize}}^h(q_0)$ is the total expected regret from period $T+1$ to n (optimization phase).

For the learning phase, the revenue from the heuristic is always nonnegative, so

$$\Delta_{i,\text{learn}}^h \leq T r^*(e_i) \leq (\frac{1}{\lambda} \log n + 1) r^*(e_i), \tag{10}$$

where $r^*(e_i)$ is the maximum revenue per period if hypothesis i is known to be true. For the optimization phase, by the definition of heuristic policy h, we have

$$\begin{aligned}
\Delta_{i,\text{optimize}}^h \leq & \mathbb{P}_i^h \left(\bigcap_{\substack{j=1 \\ j \neq i}}^{k} \{q_T^j < \frac{1}{k}\} \right) \cdot 0 + \mathbb{P}_i^h \left(\bigcup_{\substack{j=1 \\ j \neq i}}^{k} \{q_T^j \geq \frac{1}{k}\} \right) \cdot (n-T) r^*(e_i) \\
\leq & 0 + \mathbb{P}_i^h (1 - q_T^i \geq \frac{1}{k}) \cdot n r^*(e_i) \\
\leq & \frac{\mathbb{E}_i^h (1 - q_T^i)}{1/k} n r^*(e_i) \\
\leq & \frac{(2k/q_0^i) \exp(-\lambda T)}{1/k} n r^*(e_i) \\
\leq & (2k^2/q_0^i) r^*(e_i)
\end{aligned} \tag{11}$$

where the second inequality holds by Markov's inequality, the third inequality holds by Lemma 1, and the last inequality holds by definition of T.

Combining Equations (10) and (11), we have $\Delta_{i,n}^h \leq (\frac{1}{\lambda} \log n + 1 + 2k^2/q_0^i) r^*(e_i)$, so the

expected regret is bounded by

$$\Delta_n^h(q_0) = \sum_{i=1}^{k} q_0^i \Delta_{i,n}^h \le \sum_{i=1}^{k} q_0^i (\frac{1}{\lambda}\log n + 1 + 2k^2/q_0^i) r^*(e_i) \le (\frac{1}{\lambda}\log n + 1 + 2k^2) \sum_{i=1}^{k} r^*(e_i) \sim O(\log n).$$

□

The result in the Theorem 1 is favorable to a seller who uses demand learning: as the number of customers n increases, the expected regret increases in order $O(\log n)$. Comparing it to the expression in equation (7), we see that the expected revenues for both the seller and the clairvoyant increase linearly with n. Therefore, when n is large, the regret gap between the seller and the clairvoyant is relatively small.

The proof of Theorem 1 shows that for heuristic policy h, the majority of the regret is attributed to the learning phase (with order $O(\log n)$), while the regret of the optimization phase is bounded by a constant. This motivates us to do more price experimentation for the learning phase, if more price changes are allowed.

Consider the following three-price heuristic. We assume for any hypothesis i, its optimal price $p^*(e_i)$ is discriminative. For each $p^*(e_i)$, let λ_i be the constant given by Lemma 1, and let $\lambda = \min_{i=1}^{k} \lambda_i$.

Heuristic h': Let $T(n) = \lfloor \frac{1}{\lambda}\log n \rfloor + 1$. Choose any $i = 1, \ldots, k$ and apply unit price $p^*(e_i)$ for $T(T(n))$ periods. Upon learning $q_{T(T(n))}$, use the same rule as in heuristic h to determine a new price for periods $T(T(n)) + 1$ to $T(n)$. Then at period $T(n) + 1$ change price again using the same rule.

The three-price heuristic h' divides the sales horizon into three parts: one learning phase (period 1 to $T(T(n))$) and two optimization phases (period $T(T(n)) + 1$ to $T(n)$, then $T(n) + 1$ to n). Because the order of regret is determined by the length of the learning phase, it is straightforward to show that with heuristic h', the regret is of order $O(T(T(n))) = O(\log\log n)$. The more detailed proof is given in the appendix. In general, we have the following result.

Theorem 2. *Suppose the optimal price for each demand function $\varphi_i(p)$ is discriminative. If the seller is allowed to change price m times over of the sales horizon, there exists two constant C and n_0, such that for any $q_0 \in Q$ and $n \ge n_0$, $\Delta_n^*(q_0) \le C \log\log\cdots\log n$ (with m iterated logarithms).*

The matching lower bound in Cheung and Simchi-Levi (2014) shows the above upper bound is tight.

Theorem 3 (Cheung and Simchi-Levi (2014)). *Suppose the price set $\mathcal{P} = [\underline{p}, \bar{p}]$ and each demand function φ_i is continuous with a range in $(0, 1)$. Also suppose that no price in $\mathcal{P}$ achieves optimal simultaneously for all demand functions and $q_0 \neq e_i$. Then for any deterministic or random policy with m price changes,*

$$\Delta_n^*(q_0) \geq c \log^{(m)}(n),$$

where the constant $c > 0$ can depend on the demands and the policy but not on n.

5 Extended Model: Limited Inventory, Random Customer Arrival

5.1 Additional Assumptions for the Extended Model

The results of the basic model can be generalized to allow for any finite inventory level and random customer arrival process. For technical reasons, some assumptions are needed to make the problem tractable in the general setting.

In this section, we suppose the set of all allowable prices is $\mathcal{P} = [\underline{p}, \bar{p}] \cup \{p_\infty\}$ or $[\underline{p}, +\infty) \cup \{p_\infty\}$. $\underline{p}$ is the lower bound on price, and $\bar{p}$ is the upper bound (satisfying some condition that will be specified later). p_∞ is the price to "shut off" the demand. When the seller runs out of inventory, price is switched to p_∞, which implies that demand goes down to zero and the sales event is closed; we don't count that as a price change.

For all candidate demand functions $\varphi_i(p)$, $i = 1, \cdots, k$, we assume $\varphi_i(p)$ is strictly decreasing in p so that its inverse function exists. The inverse function $p_i(\varphi)$ takes value in either $[0, \bar{\varphi}]$ or $\{0\} \cup [\underline{\varphi}, \bar{\varphi}]$. As commonly assumed in economics literature, we suppose the revenue $p_i(\varphi)\varphi$ is concave in demand φ.

To show how the regret function grows when the number of periods n increases, we introduce an asymptotic regime similar to the one used by Besbes and Zeevi (2009). For any n, we suppose that the initial inventory level is $s = n\bar{s}$, where $\bar{s}$ is a constant represents inventory scarcity. Note that the total number of customer arrivals has mean ny. Thus, the index n represents the order of magnitude of both inventory and the number of customer arrivals.

5.2 Effectiveness of Limited Price Experimentation

Recall that in the basic model, a clairvoyant seller with perfect information of demand function would offer the same price for every period to maximize revenue. However, when inventory is

limited and customer arrival is random, a clairvoyant can have the incentive to use dynamic pricing. Dynamic pricing is not used here for demand learning, but for "compensating for statistical fluctuating in demand" (Gallego and Van Ryzin 1994). The expected revenue under this dynamic pricing setting is difficult to compute exactly. Therefore, we use the fixed-price policy proposed by Gallego and Van Ryzin (1994) to derive an *approximation* for the revenue. Combining the approximation technique with the results from the basic model, we will show that limited price experimentation is still very effective in improving revenue under the general setting.

Suppose demand hypothesis i holds. Using conditional expectation, it can be shown that the expected revenue in equation (5) is equal to

$$\mathbb{E}_i^\pi\left[\sum_{t=1}^n p_t X_t Y_t\right] = \mathbb{E}_i^\pi\left[\mathbb{E}_i^\pi\left[\sum_{t=1}^n p_t X_t Y_t \mid p_t\right]\right] = \mathbb{E}_i^\pi\left[\sum_{t=1}^n p_t \varphi_i(p_t) y\right] = \mathbb{E}_i^\pi\left[\sum_{t=1}^n p_i(\varphi_t)\varphi_t y\right]. \tag{12}$$

Note that although random variables X_t and Y_t do not appear in equation (12), p_t and φ_t depend implicitly on them. The last equality holds by using the inverse demand function of φ_i.

Given s units of inventory and n periods, let $R_n^\pi(s)$ be the seller's expected revenue under policy π. So the seller's maximum expected revenue is

$$\begin{aligned} R_{i,n}^*(s) &= \max_{\pi\in\Pi} R_{i,n}^\pi(s) = \max_{\pi\in\Pi} \mathbb{E}_i^\pi\left[\sum_{t=1}^n p_i(\varphi_t)\varphi_t y\right] \\ \text{subject to } &\sum_{t=1}^n X_t Y_t \le s \ \text{ a.s.} \end{aligned} \tag{13}$$

where Π is the set of pricing policies.

We also consider the following deterministic revenue maximizing problem: the customer arrival Y_t is replaced by its mean y, and demand X_t is replaced by the mean φ_t.

$$\begin{aligned} R_{i,n}^D(s) &= \max_{\varphi} \sum_{t=1}^n p_i(\varphi_t)\varphi_t y \\ \text{subject to } &\sum_{t=1}^n \varphi_t y \le s, \\ &\varphi_t \in [0,\bar{\varphi}] \quad \text{for all } t = 1,\cdots,n. \end{aligned} \tag{14}$$

The solution of the deterministic problem has the following simple characterization. Let φ^0 be the inventory run-out demand rate $\varphi^0 = \frac{s}{ny} = \frac{\bar{s}}{y}$, and let φ_i^* be the revenue maximizing rate $\varphi_i^* = \arg\max_{\varphi\in[0,\bar{\varphi}]} \varphi p_i(\varphi)$. Then the optimal solution of problem (14) is given by $\varphi_i^D =$

$\min\{\varphi^0, \varphi_i^*\}$. In the case where $\varphi_t \in \{0\} \cup [\underline{\varphi}, \bar{\varphi}]$, we assume $\underline{\varphi} < \varphi_i^D$ for all i, so the solution still holds. Motivated by the solution, we define the fixed policy F as offering $p_i(\varphi^D)$ when inventory level is positive. The expected revenue under policy F is denoted by $R_{i,n}^F(s)$.

Lemma 2. *Given time periods n and inventory level $s = n\bar{s}$, it holds that $R_{i,n}^F(s) \le R_{i,n}^*(s) \le R_{i,n}^D(s)$ for all demand hypothesis $i = 1, \cdots, k$. In addition, if $\varphi_i^* \ge \varphi^0$, there exists some constant C_1 such that $R_{i,n}^F(s) \ge R_{i,n}^D(s)(1 - \frac{C_1}{\sqrt{n}})$. If $\varphi_i^* < \varphi^0$, there exists some constant C_2 such that $R_{i,n}^F(s) \ge R_{i,n}^D(s)(1 - \frac{C_2}{n})$.*

Lemma 2 says that when the true demand is known, the revenue for the optimal policy, though difficult to compute, can be well approximated by solving simple problem (14). It also implies that the revenue of the fixed price policy F is close to the optimal revenue for large n. We therefore consider the following heuristic pricing policy, which is based on the simple heuristic h in Section 4. With this policy, the seller changes price at most once.

Heuristic h'': Choose a discriminative initial price p and apply it for $T' = (\lfloor \frac{y+\lambda+\sqrt{2y\lambda+\lambda^2}}{y^2\lambda}(\log n + 1)\rfloor + 1)$ periods, where y is the arrival rate and λ is the constant given by Lemma 1. After T' periods, choose any hypothesis i that has belief probability $q_T^i \ge 1/k$. Then apply fixed price $p_i(\varphi^D)$ before the inventory runs out. $p_i(\cdot)$ is the inverse of demand function φ_i, and φ^D is the optimal solution of (14).

Similar to Theorem 1 of the basic case, heuristic h'' can be used to bound the regret. Let $\Delta_n^*(q_0, s)$ be the minimum regret when one price change is allowed, then we have the following result.

Theorem 4. *Suppose the price set $\mathcal{P}$ contains a disriminative price, and only one price change is allowed during the sales event.*

(1) If $\varphi_i^* \ge \varphi^0$ for some hypothesis i in problem (14), there exists some constants C_1 and n_1 such that for any $n \ge n_1$ and any belief probability $q_0 \in Q$, the regret $\Delta_n^*(q_0, s) \le \Delta_n^{h''}(q_0, s) \le C_1\sqrt{n}$.

(2) If $\varphi^* < \varphi^0$ for all hypotheses $i = 1, \cdots, n$, there exists some constants C_2 and n_2 such that for any $n \ge n_2$ and any belief probability $q_0 \in Q$, the regret $\Delta_n^*(q_0, s) \le \Delta_n^{h''}(q_0, s) \le C_2 \log n$.

According to Theorem 4, if inventory is sufficient, i.e. the revenue maximizing demand is less than the inventory run-out demand ($\varphi_i^* < \varphi^0$), the regret has the same order $O(\log n)$ as in the basic model with unlimited inventory. If inventory is not sufficient ($\varphi_i^* < \varphi^0$), the regret is of

order $O(\sqrt{n})$. The regret increases because in such case, the clairvoyant can also use dynamic pricing for revenue maximization. Since the number of price changes is limited, the seller with imperfect demand information must dedicate some price changes for demand learning, and thus can't use as many prices as the clairvoyant does for revenue maximization.

The regret bound in Theorem 4 holds also for $m \geq 2$ price changes. The proof needs no modification: the seller can just use one out of m given chances. But unlike the basic case in Section 4, we are not able to get a sharper bound here for $m \geq 2$.

6 Numerical Results

In this section, we present the results of numerical experiments to analyze three aspects of our model that have not been addressed by the previous theoretical analysis. Throughout this section, we assume the latent demand function belongs to a binary set of linear demand functions $\{\varphi_1 = 1 - 0.25p, \varphi_2 = 1 - 0.5p\}$. For convenience, we let q be the probability that φ_1 is the latent demand function, so we don't need a vector for q.

6.0.1 Selecting Initial Price

The DP algorithm described in Section **??** treats the initial price as given. To find the optimal initial price, one can solve the DP using different initial prices, but that would require an enormous amount of computation. However, for the approximate DP algorithm, the initial price to maximize the approximated revenue-to-go function in Equation (**??**) is the "greedy price" $p^*(q_0)$, the price that maximizes revenue for the first period.

We test the effectiveness of this greedy rule by using $q = 0.25, 0.5, 0.75$ and the corresponding greedy prices $p^*(q) = 1.14, 1.33, 1.60$. We also include a fourth price $p = 0.8$ as a comparison. The fixed price policy uses the greedy price for the entire sales horizon. The revenue loss is defined as the expected regret (under a given initial price) as a percentage of a clairvoyant's revenue.

Figure 1 shows the revenue loss under different initial prices for $n = 10$ customers. The result indicates that the greedy rule provides a good estimation but it does not always hold. When $q_0 = 0.25$ and $q_0 = 0.75$, the greedy price has the smallest revenue loss. When $q_0 = 0.5$, the greedy price $p = 1.33$ is slightly worse than $p = 1.6$. The non-greedy price $p = 0.8$ is so bad that its curve overlaps the fixed price curve.

Figure 2 shows the revenue loss for $n = 100$ customers. The result shows the revenue loss is much smaller in this case. For all three greedy prices, the revenue loss is within 4%. The figure

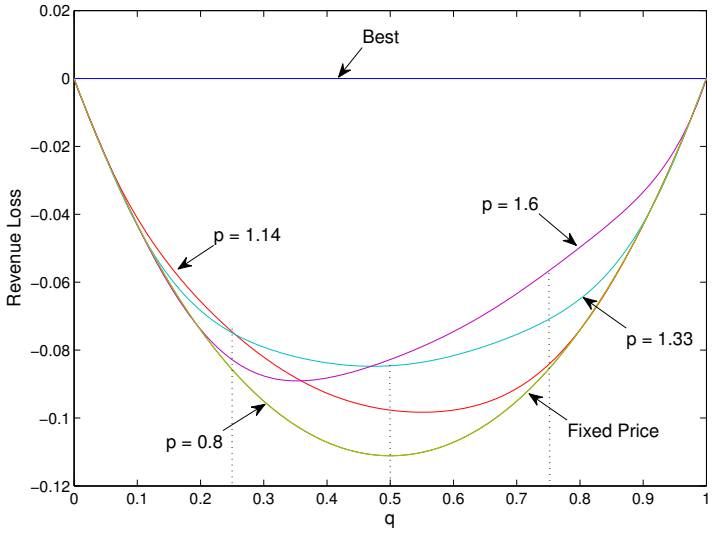

Figure 1: Revenue loss for $n = 10$.

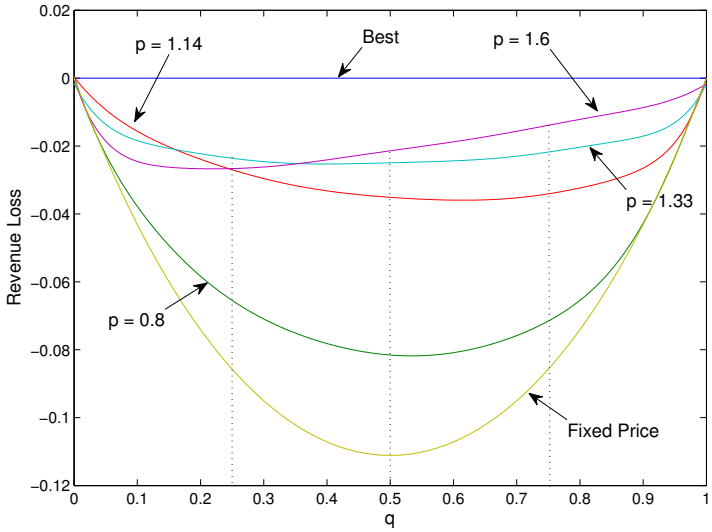

Figure 2: Revenue loss for $n = 100$.

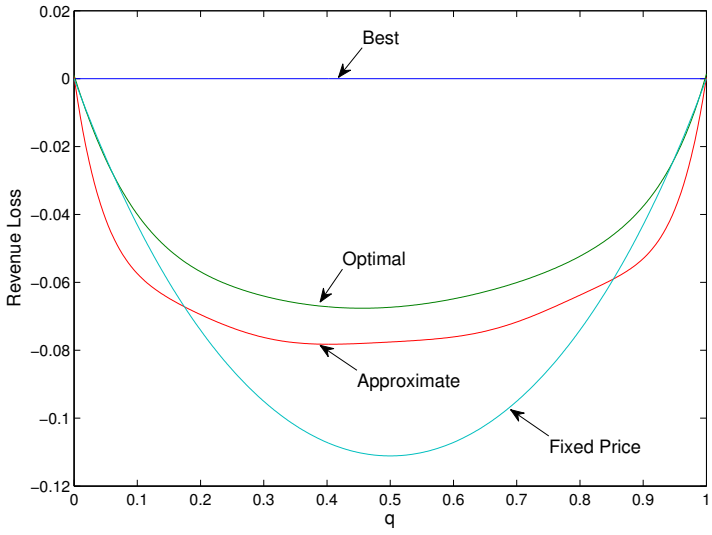

Figure 3: Approximate DP algorithm for $n = 20$.

also indicates that the greedy rule does not hold. For $q = 0.25$, the corresponding greedy price $p = 1.14$ is dominated by price $p = 1.33$; for $q = 0.5$, the greedy price $p = 1.33$ is dominated by $p = 1.6$. This implies that a higher price is usually more favorable than the greedy price because, in this set of demand functions, a higher price is more discriminative (i.e., $\varphi_1(p) - \varphi_2(p)$ is increasing in p). So even if a higher price is not the best price for exploitation, it is better for exploration.

6.0.2 Approximation Algorithm

In this part, we test the performance of the approximate DP algorithm. The initial price is set to $p = 1.33$. Figure 3 shows the result for $n = 20$ customers. The result shows the expected revenue for approximation algorithm is within 2% for all q. But when q is close to 0 and 1, it performs worse than the fixed price algorithm, because the approximation algorithm is biased on continuing. Therefore, we modify the approximation algorithm so that it is forced to stop learning when q is close to 0 and 1.[6] The performance of the modified algorithm for $n = 100$ is shown in Figure 4.

[6] In this experiment, we force the algorithm to stop learning when $q < 0.03$ or $q > 0.97$. Because the fixed price policy in those two cases has a revenue loss less than 2%, this guarantees that the modified approximation algorithm is at most 2% worse than the original one.

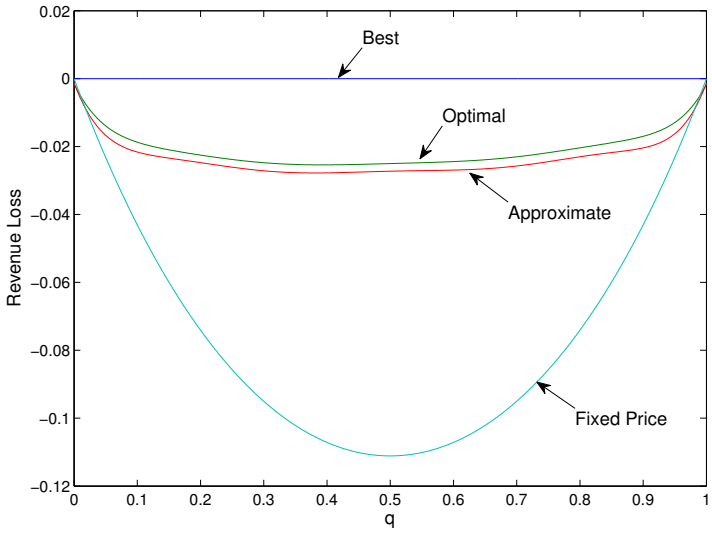

Figure 4: Modified approximate DP algorithm for $n = 100$.

6.0.3 Optimal Stopping Time

In this part, we present the results of numerical experiments to identify the average time to change price (the stopping time) under the two-price policy given by the approximate DP algorithm. The approximate DP algorithm is biased on continuing, so the stopping time under the optimal policy would be shorter than the figures we show. We sought an upper bound on the optimal stopping time by initializing the least informative prior belief probability $q = 0.5$. The initial price is $p = 1.33$. We simulated 100 sample sequences of customer purchase decisions to identify the stopping time.

The average stopping time is shown in Figure 5 (the x-coordinate is scaled by logarithm). The result shows that when the number of customers increases, the stopping time grows much slower. For example, when $n = 10000$, the stopping time is $T = 49.16$, meaning the seller spends only a fraction of time on demand learning.

7 Conclusion

This paper considers a dynamic pricing problem where the underlying demand model is unknown and opportunities for price experimentation are limited due to business constraints. We assume the latent demand function is drawn from a known finite set. Customers arrive in a sequential fashion. The seller observes the success or failure of each sales attempt, and then updates his

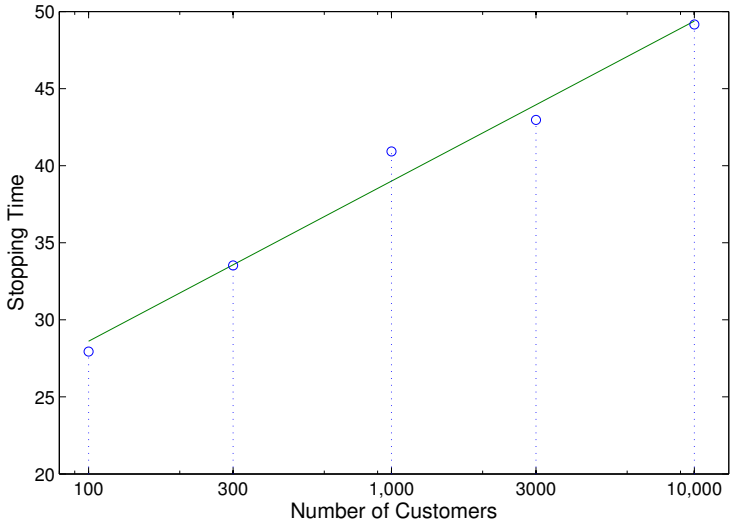

Figure 5: Average stopping time.

belief on the latent demand function. The seller can adjust the price dynamically to maximize revenue. However, during the course of the sales window, the seller can only change the price a limited number of times.

In this model setting, we find that there is only a small gap in revenue between a seller using limited price changes and a clairvoyant who has full information on the demand function. In particular, assuming customer arrivals are deterministic and inventory is unconstrained, when there are n customer arrivals during the sales horizon, the gap of regret is bounded by $O(\log n)$ if the seller changes price only once during the sales window; by changing price m times, the gap can be further reduced to $O(\log \cdots \log n)$, with m iterations of the logarithm. More generally, when customer arrival process is stochastic and inventory is constrained, the regret is bounded by $O(\sqrt{n})$ (n represents both the scale of customers and the scale inventory.) For the basic case where the seller can change price only once and inventory is unconstrained, we analyze the optimal pricing policy using dynamic programming, and develop an approximate dynamic programming algorithm.

Appendix

Proof of Lemma 1. The following proof is modified from Harrison et al. (2012), where they assume only two possible demand functions ($k = 2$). Without loss of generality, we prove the case when the true hypothesis is $i = 1$. We simplify the denotation by using $\mathbb{E}(\cdot)$ for $\mathbb{E}_1^\tau(\cdot)$, and

$\mathbb{P}(\cdot)$ for $\mathbb{P}_1^\pi(\cdot)$. By Equation (3), it is readily verified that the belief probability for hypothesis 1 at period t is

$$q_t^1 = \frac{q_0^1 \varphi_1(p)^{\sum_{j=1}^t X_j} \bar{\varphi}_1(p)^{\sum_{j=1}^t (1-X_j)}}{q_0^1 \varphi_1(p)^{\sum_{j=1}^t X_j} \bar{\varphi}_1(p)^{\sum_{j=1}^t (1-X_j)} + \cdots + q_0^k \varphi_k(p)^{\sum_{j=1}^t X_j} \bar{\varphi}_k(p)^{\sum_{j=1}^t (1-X_j)}},$$

where $\bar{\varphi}_i = 1 - \varphi_i$, $\forall i = 1, \ldots, k$. So

$$\begin{aligned}
\mathbb{E}[q_t^1] =& \mathbb{E}\left[\frac{q_0^1 \varphi_1(p)^{\sum_{j=1}^t X_j} \bar{\varphi}_1(p)^{\sum_{j=1}^t (1-X_j)}}{\sum_{i=1}^k q_0^i \varphi_i(p)^{\sum_{j=1}^t X_j} \bar{\varphi}_i(p)^{\sum_{j=1}^t (1-X_j)}}\right] \\
=& \mathbb{E}\left[\frac{1}{1 + \sum_{i=2}^k \frac{q_0^i}{q_0^1} \left(\frac{\varphi_i(p)}{\varphi_1(p)}\right)^{\sum_{j=1}^t X_j} \left(\frac{\bar{\varphi}_i(p)}{\bar{\varphi}_1(p)}\right)^{\sum_{j=1}^t (1-X_j)}}\right] \\
\geq& 1 - \sum_{i=2}^k \mathbb{E}\left[\frac{q_0^i}{q_0^1} \left(\frac{\varphi_i(p)}{\varphi_1(p)}\right)^{\sum_{j=1}^t X_j} \left(\frac{\bar{\varphi}_i(p)}{\bar{\varphi}_1(p)}\right)^{\sum_{j=1}^t (1-X_j)}\right] \\
=& 1 - \sum_{i=2}^k \mathbb{E}\left[\frac{q_0^i}{q_0^1} \exp\left(\sum_{j=1}^t X_j \log \frac{\varphi_i(p)}{\varphi_1(p)} + \sum_{j=1}^t (1-X_j) \log \frac{\bar{\varphi}_i(p)}{\bar{\varphi}_1(p)}\right)\right] \\
=& 1 - \sum_{i=2}^k \mathbb{E}\left[\frac{q_0^i}{q_0^1} \exp\left(t\Big[\varphi_1(p) \log \frac{\varphi_i(p)}{\varphi_1(p)} + \bar{\varphi}_1(p) \log \frac{\bar{\varphi}_i(p)}{\bar{\varphi}_1(p)}\Big] + \Big(\sum_{j=1}^t X_j - \varphi_1(p)t\Big) \log \frac{\varphi_i(p)\bar{\varphi}_1(p)}{\varphi_1(p)\bar{\varphi}_i(p)}\right)\right].
\end{aligned} \tag{15}$$

For brevity, we denote by B_i the expression in each expectation.

Note that $\log(\cdot)$ is strictly concave and $\varphi_1(p) \neq \varphi_i(p)$. So for any i, it holds that

$$\delta_i := -\left[\varphi_1(p) \log \frac{\varphi_i(p)}{\varphi_1(p)} + \bar{\varphi}_1(p) \log \frac{\bar{\varphi}_i(p)}{\bar{\varphi}_1(p)}\right] > 0.$$

Let $\epsilon_i > 0$ be a constant such that

$$-\left[\varphi_1(p) \log \frac{\varphi_i(p)}{\varphi_1(p)} + \bar{\varphi}_1(p) \log \frac{\bar{\varphi}_i(p)}{\bar{\varphi}_1(p)}\right] - \epsilon_i \left|\log \frac{\varphi_i(p)\bar{\varphi}_1(p)}{\varphi_1(p)\bar{\varphi}_i(p)}\right| > \frac{\delta_i}{2}.$$

Because $|X_j| \leq 1$, by Hoeffding's inequality, we have

$$\mathbb{P}\Big(\Big|\sum_{j=1}^t X_j - \varphi_1(p)t\Big| \geq xt\Big) \leq 2\exp(-tx^2/2).$$

Define events $A_i := \{\big|\sum_{j=1}^t X_j - \varphi_1(p)t\big| \geq \epsilon_i t\}$ for $i \geq 2$, so $\mathbb{P}(A_i) \leq 2\exp(-t\epsilon_i^2/2)$. By

Equation (15), we have $\mathbb{E}[B_i \mid A_i^c] \leq (q_0^i/q_0^1)\exp(-t\delta_i/2)$. Thus,

$$\begin{aligned}
\mathbb{E}[q_t^1] =& \mathbb{E}[q_t^1 \mid \cap_{i=2}^k A_i^c] \cdot \mathbb{P}(\cap_{i=2}^k A_i^c) + \mathbb{E}[q_t^1 \mid \cup_{i=2}^k A_i] \cdot \mathbb{P}(\cup_{i=2}^k A_i) \\
\geq& \mathbb{E}[q_t^1 \mid \cap_{i=2}^k A_i^c] \cdot \mathbb{P}(\cap_{i=2}^k A_i^c) \\
=& (1 - \sum_{i=2}^k \mathbb{E}[B_i \mid \cap_{i=2}^k A_i^c])(1 - \mathbb{P}(\cup_{i=2}^k A_i)) \\
\geq& 1 - \sum_{i=2}^k \mathbb{E}[B_i \mid \cap_{i=2}^k A_i^c] - \mathbb{P}(\cup_{i=2}^k A_i) \\
\geq& 1 - \sum_{i=2}^k \mathbb{E}[B_i \mid \cap_{i=2}^k A_i^c] - \sum_{i=2}^k \mathbb{P}(A_i) \\
\geq& 1 - \sum_{i=2}^k \frac{q_0^i}{q_0^1}\exp(-t\delta_i/2) - \sum_{i=2}^k 2\exp(-t\epsilon_i^2/2) \\
\geq& 1 - \frac{1-q_0^1}{q_0^1}\exp(-\lambda t) - 2(k-1)\exp(-\lambda t) \\
\geq& 1 - \frac{2k}{q_0^1}\exp(-\lambda t),
\end{aligned}$$

where $\lambda := \min_{i=2}^k \{\delta_i/2, \epsilon_i^2/2\}$.

Now that a constant λ exists for each case $i = 1, \ldots,$ we simplify take the smallest λ over all cases so these constants are universal. □

Sketch Proof of Theorem 2. We only prove the theorem for $m = 2$, because the proof for the general case is similar but requires more involved notation. For two price changes, using heuristic h', the regret for hypothesis i is:

$$\Delta_{i,n}^{h'} = \Delta_{i,1,T(T(n))}^{h'} + \Delta_{i,T(T(n))+1,T(n)}^{h'} + \Delta_{i,T(n)+1,n}^{h'},$$

where the first term is the regret for the first phase (period 1 to period $T(T(n))$), and so forth. There are $T(T(n))$ periods for the first phase, so

$$\Delta_{i,1,T(T(n))}^{h'} \sim O(T(T(n))) = O(\log\log n).$$

The second price is used from period $T(T(n)) + 1$ to period $T(n)$. According to the proof of Theorem 1, this can be viewed as an optimization phase, so the regret for the second phase $\Delta_{i,T(T(n))+1,T(n)}^{h'}$ is bound by $O(1)$.

The seller keeps learning throughout the first and the second phase. We can view the first and the second phases jointly as one learning phase. Therefore, by using Theorem 1 again,

the regret for the last phase $\Delta^{h'}_{i,T(n)+1,n}$ is bounded by $O(1)$. (Lemma 1 is proved under a single price but it also holds for a finite set of discriminative prices. Alternatively, one can use $T(T(n)) + T(n)$ as the time for the second price change, so Lemma 1 is still applicable.)

Combining the regret bound for the three phases, we get $\Delta^{h'}_{i,n} \sim O(\log\log n)$. Taking the weighted average over q_0, we have $\Delta^*_n(q_0) \le \Delta^{h'}_n(q_0) \sim O(\log\log n)$. □

Proof of Lemma 2. The proof follows Theorem 2 and 3 of Gallego and Van Ryzin (1994), although minor changes are made because the original proof is for a continuous time model.

Note that $R^F_{i,n}(s) \le R^*_{i,n}(s)$ by definition. To show the second inequality, we will prove $R^\pi_{i,n}(s) \le R^D_{i,n}(s)$ for any policy π, and thus $R^*_{i,n}(s) = \sup_{\pi\in\Pi} R^\pi_{i,n}(s) \le R^D_{i,n}(s)$ holds.

We first consider the case where $\varphi \in [0, \bar\varphi]$. For $\theta \ge 0$, we define the augmented revenue function for the stochastic problem (13) by

$$R^\pi_{i,n}(s,\theta) = \mathbb{E}^\pi_i\Big[\sum_{t=1}^n p_t X_t Y_t\Big] - \theta\left(\mathbb{E}^\pi_i\Big[\sum_{t=1}^n X_t Y_t\Big] - s\right) = \mathbb{E}^\pi_i\Big[\sum_{t=1}^n (p_i(\varphi_t) - \theta)\varphi_t y\Big] + \theta s.$$

Because $\sum_{t=1}^n p_t X_t Y_t \le s$ almost surely, we have $R^\pi_{i,n}(s) \le R^\pi_{i,n}(s,\theta)$ for all $\theta \ge 0$. Similarly, we define the augmented revenue function for the deterministic problem (14) by

$$R^D_{i,n}(s,\theta) = \max_{\varphi} \sum_{t=1}^n (p_i(\varphi_t) - \theta)\varphi_t y + \theta s.$$

It holds that

$$\begin{aligned} R^\pi_{i,n}(s,\theta) &= \sum_{t=1}^n \mathbb{E}^\pi_i\big[(p_i(\varphi_t) - \theta)\varphi_t y\big] + \theta s \\ &\le \sum_{t=1}^n \left(\max_{\varphi_t\in[0,\bar\varphi]} (p_i(\varphi) - \theta)\varphi_t y\right) + \theta s \\ &= \max_{\varphi_t\in[0,\bar\varphi]} \sum_{t=1}^n (p_i(\varphi - t) - \theta)\varphi_t y + \theta s = R^D_{i,n}(s,\theta), \end{aligned}$$

and thus we have

$$R^\pi_{i,n}(s) \le \inf_{\theta\ge 0} R^\pi_{i,n}(s) \le \inf_{\theta\ge 0} R^D_{i,n}(s,\theta).$$

Note that problem (14) is a convex programming problem with Slater point $\varphi_t = 0$, so by strong duality, $\inf_{\theta\ge 0} R^D_{i,n}(s,\theta) = R^D_{i,n}(s)$.

In the case where $\varphi \in \{0\} \cup [\underline\varphi, \bar\varphi]$, the above argument again implies $R^\pi_{i,n}(s)$ (subject to $\varphi_t \in \{0\} \cup [\underline\varphi, \bar\varphi]$) $\le R^D_{i,n}(s)$ (subject to $\varphi_t \in [0, \bar\varphi]$). By assumption, $R^D_{i,n}(s)$ (subject to $\varphi_t \in [0, \bar\varphi]$) $= R^D_{i,n}(s)$ (subject to $\varphi_t \in \{0\} \cup [\underline\varphi, \bar\varphi]$), so the inequality still holds.

We then prove the second part of the lemma. Let $S = \sum_{t=1}^{n} X_t Y_t$ be the number of units sold, with mean μ and standard deviation σ. According to Gallego and Van Ryzin (1994), it holds that

$$\mathbb{E}[(S-s)^+] \leq \frac{\sqrt{\sigma^2 + (s-\mu)^2} - (s-\mu)}{2}.$$

Since $\varphi^D = \min\{\varphi^0, \varphi_i^*\}$, we consider two cases. If $\varphi_i^* \geq \varphi^0$, we have $\mathbb{E}[S] = s = n\bar{s}$, $\mathrm{Var}(S) = s(1-s/n) = n\bar{s}(1-\bar{s})$. So

$$R_{i,n}^F(s) = p_i(\varphi^0)\mathbb{E}_i^F[S-(S-s)^+] \geq p_i(\varphi^0)\left[s - \frac{\sqrt{s(1-s/n)}}{2}\right] = R_{i,n}^D(s)\left[1 - \frac{\sqrt{(1-\bar{s})/\bar{s}}}{2\sqrt{n}}\right].$$

If $\varphi_i^* < \varphi^0$, we have $\mathbb{E}[S] = ny\varphi_i^*$, $\mathrm{Var}(S) = ny\varphi_i^*(1-\varphi_i^*)$, and similarly

$$\begin{aligned} R_{i,n}^F(s) = p_i(\varphi_i^*)\mathbb{E}_i^F[S-(S-s)^+] &\leq p_i(\varphi_i^*)\left[ny\varphi_i^* - \frac{\sqrt{ny\varphi_i^*(1-y\varphi_i^*) + n^2(\bar{s}-y\varphi_i^*)^2} - n(\bar{s}-y\varphi_i^*)}{2}\right] \\ &\leq p_i(\varphi_i^*)\left[ny\varphi_i^* - \frac{ny\varphi_i^*(1-y\varphi_i^*)}{4n(\bar{s}-y\varphi_i^*)}\right] = R_{i,n}^F(s)\left[1 - \frac{1-y\varphi_i^*}{4n(\bar{s}-y\varphi_i^*)}\right] \end{aligned}$$

□

Proof of Theorem 4. Suppose hypothesis i is true, consider the regret:

$$\Delta_{i,n}^{h''}(s) = \tilde{R}_{i,n}(s) - R_{i,n}^{h''}(s),$$

where $\tilde{R}_{i,n}(s)$ is the clairvoyant's revenue and $R_{i,n}^{h''}(s)$ is the revenue for heuristic h''. By Lemma 2, $\tilde{R}_{i,n}(s) \leq R_{i,n}^D(s)$, so

$$\Delta_{i,n}^{h''}(s) \leq R_{i,n}^D(s) - R_{i,n}^{h''}(s).$$

Note that heuristic h'' divides the sales horizon into a learning phase (period 1 to T') and an optimization phase (peirod $T'+1$ to n). We first show that in the learning phase, heuristic h'' guarantees sufficient demand learning with high probability.

Lemma 3. *Let $T = \lfloor \frac{1}{\lambda}\log n \rfloor + 1$, $c = (y + \lambda + \sqrt{2y\lambda + \lambda^2})/y^2$ and $T' = \lfloor cT \rfloor + 1$. For any policy π and demand hypothesis i,*

$$\mathbb{P}_i^\pi\left(\sum_{t=1}^{T'} Y_t \leq T\right) \leq \frac{1}{n}.$$

Proof of Lemma 3. Since Y_t's are independent Bernoulli random variables with mean y, $M_n = \sum_{t=1}^{n} Y_t - yn$ is a martingale that satisfies $|M_n - M_{n-1}| \leq 1$. By Hoeffding's inequality,

$$\begin{aligned}\mathbb{P}_i^{\pi}\left(\sum_{t=1}^{T'} Y_t \leq T\right) &= \mathbb{P}_i^{\pi}\left(\sum_{t=1}^{T'} Y_t - yT' \leq T - yT'\right) \leq \exp\left(-\frac{(T-yT')^2}{2T'}\right) \\ &\leq \exp\left(-\frac{T^2(1-yc)^2}{2cT}\right) \leq \exp\left(-\frac{\log n/\lambda \cdot (1-yc)^2}{2c}\right) = \frac{1}{n}\end{aligned}$$

□

Let A_i be the event that the seller chooses hypothesis i after period T'. By Lemma 3 with Theorem 1, we have

$$\begin{aligned}\mathbb{P}_i^{h''}(A_i^c) \leq& \mathbb{P}_i^{h''}(\{\text{learning sample} < T\} \cup \{\text{select the wrong model with } T \text{ samples}\}) \\ \leq& \mathbb{P}_i^{h''}(\{\text{learning sample} < T\}) + \mathbb{P}_i^{h''}(\{\text{select the wrong model with } T \text{ samples}\}) \\ \leq& \frac{1}{n} + \frac{2k^2}{q_0^i n}.\end{aligned}$$

Therefore, the revenue for heuristic h'' is bounded below by

$$R_{i,n}^{h''}(s) \geq \mathbb{E}_i^{h''}\left[\sum_{i=1}^{n} X_t Y_t p_t \mid A_i\right]\mathbb{P}(A_i) \geq \mathbb{E}_i^{h''}\left[\sum_{i=1}^{n} X_t Y_t p_t \mid A_i\right] \cdot (1 - \frac{2k^2}{q_0^i n} - \frac{1}{n}).$$

Finally, we have

$$\mathbb{E}_i^{h''}\left[\sum_{i=1}^{n} X_t Y_t p_t \mid A_i\right] \geq R_{i,n}^{F}(s) - p_i^D T',$$

where p_i^D is the fixed price offered in policy F. To see why this is true, note that heuristic policy h'' and fixed price policy F offer the same price if event A_i happens, expect for the learning phase. Since at most T' items can be sold during the learning phase, the difference in revenue is at most $p_i^D T'$. In sum, we have

$$\begin{aligned}\Delta_{i,n}^{h''}(s) \leq& R_{i,n}^{D}(s) - R_{i,n}^{h''}(s) \\ \leq& R_{i,n}^{D}(s) - (R_{i,n}^{F}(s) - p_i^D T') \cdot (1 - \frac{2k^2}{q_0^i n} - \frac{1}{n}) \\ \leq& R_{i,n}^{D}(s) - R_{i,n}^{F}(s) + R_{i,n}^{F}(s) \cdot (\frac{2k^2}{q_0^i n} + \frac{1}{n}) + p_i^D T' \\ \leq& R_{i,n}^{D}(s) - R_{i,n}^{F}(s) + r^*(e_i) yn \cdot (\frac{2k^2}{q_0^i n} + \frac{1}{n}) + p_i^D (\lfloor \frac{y + \lambda + \sqrt{2y\lambda + \lambda^2}}{y^2 \lambda}(\log n + 1) \rfloor + 1) \\ \leq& R_{i,n}^{D}(s) - R_{i,n}^{F}(s) + C^i \log n,\end{aligned}$$

where C^i is a constant large enough so that the last step holds.

By Lemma 2, if $\varphi_i^* \geq \varphi^0$ for some hypothesis

$$\begin{aligned}
\Delta_n^{h''}(q_0, s) =& \sum_{i=1}^{n} q_0^i \Delta_{i,n}^{h''}(s) \\
\leq& \sum_{i=1}^{n} q_0^i \left[R_{i,n}^D(s) - R_{i,n}^F(s) + C^i \log n \right] \\
\leq& \sum_{i=1}^{n} q_0^i \left[R_{i,n}^D(s) \max\{C_1^i, C_2^i\}/\sqrt{n} + C^i \log n \right] \\
\leq& \sum_{i=1}^{n} q_0^i \left[\varphi_i^D \max\{C_1^i, C_2^i\}\sqrt{n} + C^i \log n \right] \\
\leq& C_1\sqrt{n} + C \log n,
\end{aligned}$$

where C_1 and C are some constants large enough so the last step holds. Otherwise, if $\varphi_i^* < \varphi^0$ for all hypotheses, we have

$$\begin{aligned}
\Delta_n^{h''}(q_0, s) =& \sum_{i=1}^{n} q_0^i \Delta_{i,n}^{h''}(s) \\
\leq& \sum_{i=1}^{n} q_0^i \left[R_{i,n}^D(s) - R_{i,n}^F(s) + C^i \log n \right] \\
\leq& \sum_{i=1}^{n} q_0^i \left[R_{i,n}^D(s) C_2^i/n + C^i \log n \right] \\
\leq& \sum_{i=1}^{n} q_0^i \left[\varphi_i^D C_2^i + C^i \log n \right] \\
\leq& C_2 + C \log n,
\end{aligned}$$

where C_2 and C are some constants large enough so the last step holds. □

References

Araman, V. F. and Caldentey, R. (2009). Dynamic pricing for nonperishable products with demand learning. *Operations research*, 57(5):1169–1188.

Aviv, Y., Levin, Y., and Nediak, M. (2009). Counteracting strategic consumer behavior in dynamic pricing systems. In Tang, C. S. and Netessine, S., editors, *Consumer-Driven Demand and Operations Management Models*, volume 131 of *International Series in Operations Research & Management Science*, pages 323–352. Springer US.

Aviv, Y. and Pazgal, A. (2005). Dynamic pricing of short life-cycle products through active learning. working paper, Olin School Business, Washington Univ., St. Louis, MO.

Aviv, Y. and Vulcano, G. (2012). Dynamic list pricing. In Özer, Ö. and Phillips, R., editors, *The Oxford Handbook of Pricing Management*, pages 522–584. Oxford University Press, Oxford.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. (2013). Bandits with knapsacks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 207–216. IEEE.

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

Besbes, O. and Zeevi, A. (2013). On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Columbia Business School Research Paper*, (12-5).

Bitran, G. R. and Mondschein, S. V. (1997). Periodic pricing of seasonal products in retailing. *Management Science*, 43(1):64–79.

Broder, J. (2011). *Online Algorithms For Revenue Management*. PhD thesis, Cornell University.

Broder, J. and Rusmevichientong, P. (2012). Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980.

Caro, F. and Gallien, J. (2012). Clearance pricing optimization for a fast-fashion retailer. *Operations Research*, 60(6):1404–1422.

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

Cheung, W. C. and Simchi-Levi, D. (2014). Lower bounds for dynamic pricing with limited experimentation. Working paper, Massachusetts Institute of Technology, Cambridge, MA.

Cope, E. (2007). Bayesian strategies for dynamic pricing in e-commerce. *Naval Research Logistics*, 54(3):265–281.

den Boer, A. (2014a). Dynamic pricing and learning: Historical origins, current research, and new directions. Working paper, University of Twente, Netherland.

den Boer, A. (2014b). Dynamic pricing with multiple products and partially specified demand distribution. *Mathematics of Operations Research*. To appear.

den Boer, A. and Zwart, B. (2014). Simultaneously learning and optimizing using controlled variance pricing. *Management Science*, 60(3):770–783.

Farias, V. and Van Roy, B. (2010). Dynamic pricing with a prior on market response. *Operations Research*, 58(1):16–29.

Feng, Y. and Gallego, G. (1995). Optimal starting times for end-of-season sales and optimal stopping times for promotional fares. *Management Science*, 41(8):1371–1391.

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

Harrison, J., Keskin, N., and Zeevi, A. (2012). Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science*, 58(3):570–586.

Keskin, N. B. and Zeevi, A. (2013). Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. working paper.

Kleinberg, R. and Leighton, T. (2003). The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 594–605. IEEE.

Mersereau, A. J., Rusmevichientong, P., and Tsitsiklis, J. N. (2009). A structured multi-armed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control*, 54(12):2787–2802.

Netessine, S. (2006). Dynamic pricing of inventory/capacity with infrequent price changes. *European Journal of Operational Research*, 174(1):553–580.

Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411.

Wang, Z., Deng, S., and Ye, Y. (2014). Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research*, 62(2):318–331.

Wolverson, R. (2012). High and low: Online flash sales go beyond fashion to survive. *Time Magazine*, 180(19):Special Section 9–12.

Zbaracki, M. J., Ritson, M., Levy, D., Dutta, S., and Bergen, M. (2004). Managerial and customer costs of price adjustment: direct evidence from industrial markets. *Review of Economics and Statistics*, 86(2):514–533.