Parametric demand learning with limited price explorations in a backlog stochastic inventory system

Boxiao Chen and Xiuli Chao

ABSTRACT
We study a multi-period stochastic inventory system with backlogs. Demand in each period is random and price sensitive, but the firm has little or no prior knowledge about the demand distribution and how each customer responds to the selling price, so the firm has to learn the demand process when making periodic pricing and inventory replenishment decisions to maximize its expected total profit. We consider the scenario where the firm is faced with the business constraint that prevents it from conducting extensive price exploration, and develop parametric data-driven algorithms for pricing and inventory decisions. We measure the performances of the algorithms by regret, which is the profit loss compared with a clairvoyant who has complete information about the demand distribution. We analyze the cases where the number of price changes is restricted to a given number or a small number relative to the planning horizon, and show that the regrets for the corresponding learning algorithms converge at the best possible rates in the sense that they reach the theoretical lower bounds. Numerical results indicate that these algorithms empirically perform very well. Supplementary materials are available for this article. Go to the publisher’s online edition of IISE Transactions, for datasets, additional tables, detailed proofs, etc.

ARTICLE HISTORY
Received 5 April 2018
Accepted 24 September 2018

KEYWORDS
Backlog inventory control; parametric demand learning; limited price exploration; asymptotic optimality

1. Introduction
Information about potential demand and customers’ responses to price change is critical for making pricing and inventory decisions. However, in many applications, especially for new products, such information is not known a priori, and the firm needs to learn such information on the fly through, for example, price experimentation. In practice, however, firms may not want to conduct very frequent price experimentation, because frequent price changes can confuse customers, hurt brand reputation, and incur operational costs (e.g., the cost of changing price labels in brick-and-mortar stores). See Cheung et al. (2017) for more discussions on this issue.

In this article we consider a dynamic joint pricing and inventory control problem over a finite planning horizon, where the firm has limited knowledge about the potential demand distribution and customers’ responses to selling prices. The firm dynamically determines its inventory replenishment and pricing decisions in each period, subject to some constraint on the number of price changes, and the objective is to maximize its total expected profit. We consider the setting that the number of potential customers in each period follows a discrete, but unknown, distribution, and each customer’s response (i.e., the probability to purchase) on an offered price is drawn from a family of distributions with unknown parameters. We develop data-driven algorithms to compute pricing and inventory replenishment decisions when there exist constraints on the number of price changes, and evaluate the performances of the learning algorithms by regret, which is the total profit loss compared with a clairvoyant who has complete information about the potential demand distribution and customer response to the selling price.

Specifically, we study the scenarios where the number of price changes is limited to either no more than a positive constant, or a small number compared with the length of the planning horizon (in the order of \( \log T \) when the length of the planning horizon is \( T \)), and develop a learning algorithm for each case. We derive the regret for the learning algorithm and determine its dependency on the number of price changes allowed. For each scenario, we show that the regret of the proposed algorithm converges at the best possible rate in the sense that its regret rate matches that of the lower bound of any learning algorithms for the respective problems. We also conduct numerical studies on these algorithms and show that they empirically perform very well.

1.1. Comparisons with the literature
This article is related to the research literature dealing with limited demand information in stochastic inventory control, revenue management, and joint pricing and inventory
control problems. For each category, the research literature is classified as taking either a parametric or a non-parametric approach.

Studies on stochastic inventory models, with complete or limited information about demand process, have a long history. In the literature on stochastic inventory control with parameter estimation, early research papers include Scarf (1959, 1960), Murray and Silver (1966), Azoury (1985), and Lovejoy (1990) for completely observable demand data; Lariviere and Porteus (1999) and Ding et al. (2002) for censored demand; and Chen and Plambeck (2008) for the case with multiple products. For non-parametric approaches, when demand is not censored, Bookbinder and Lordahl (1989) use bootstrap sampling, and Levi et al. (2007) and Levi et al. (2010) estimate the true objective function (e.g., the expected cost or profit) using sample average approximation. For censored demand, Huh and Rusmevichientong (2009) and Huh et al. (2011) present algorithms based on on-line convex optimization and Kalman filtering, respectively. In these papers, price is static and exogenous, thus the firm is only concerned with inventory replenishment decisions.

In the revenue management literature, early papers such as Kalish (1983), and Gallego and van Ryzin (1994) consider a firm’s pricing problem when it has complete information about the underlying demand process. These papers have been extended to parametric settings (i.e., the firm knows the family of demand distribution, but not the parameters of the distribution) by, e.g., Aviv and Pazgal (2005), Carvalho and Puterman (2005), Araman and Caldentey (2009), Den Boer and Zwart (2010), Farias and Van Roy (2010), Broder and Rusmevichientong (2012), Harrison et al. (2012), and Keskin and Zeevi (2014), among others. Broder (2011) and Cheung et al. (2017) develop learning algorithms for a pricing problem with constraint on the number of price changes. On the other hand, Besbes and Zeevi (2009, 2015), Wang et al. (2014), Chen, Jasin and Duenyas (2018), consider non-parametric versions of pricing problems. Besbes and Zeevi (2009) propose an algorithm that first learns about the demand information then applies it in the rest of the planning horizon. The regret convergence results of Besbes and Zeevi (2009) are strengthened in Wang et al. (2014).

There are numerous studies in the literature on joint pricing and inventory decisions. As in the above two categories, early papers in this area, including Whitin (1955), Karlin and Carr (1962), Thowsen (1975), Federgruen and Heching (1999), and Chen and Simchi-Levi (2004), assume that the firm has complete information about the demand distribution. Refer to survey papers by Elmaghriby and Keskinocak (2003), Chan et al. (2004), and Chen and Simchi-Levi (2012) for more references. When complete information is not available, parametric models are studied in Subrahmanyan and Shoemaker (1996), and Petruzz and Dada (2002). Burnetas and Smith (2000) consider this problem in a non-parametric setting, and apply the multi-armed bandit approach to develop a data-driven policy. The non-parametric setting is also studied in Chen, Chao and Ahn (2018) who obtain a best possible regret convergence rate for their proposed learning algorithm.

Our work belongs to the parametric category of joint optimization of pricing and inventory control. The most closely related works to ours are Besbes and Zeevi (2009), Broder (2011), Broder and Rusmevichientong (2012), Cheung et al. (2017). Broder (2011) and Broder and Rusmevichientong (2012) consider dynamic pricing problems in which there is a single arrival in each period, and Broder (2011) further considers limited price experimentations. In our model, the number of arrivals per period is random, of which the distribution is not known and needs to be learned from data. Both Broder (2011) and Broder and Rusmevichientong (2012) assume infinite initial inventory and zero holding cost, hence there are no inventory replenishment decisions. In our model, the firm makes joint decisions for pricing and inventory control. As the inventories are non-perishable, some prescribed inventory level target (by a learning algorithm) cannot be achieved if it is lower than the carry-over inventories from the previous period. Therefore, the learning problem we study has a constraint due to carry-over inventories. See Huh and Rusmevichientong (2009) for more discussions on the impact of carry-over inventory on the performance of learning algorithms.

Cheung et al. (2017) study a dynamic pricing problem with demand learning, which also consider limited price changes. As in the aforementioned literature, they also assume initial inventory is infinite, and there is exactly one customer in each period. Therefore, they have no inventory replenishment decision. The firm does not know the demand function, but knows that it belongs to a finite set of functions that is known. In contrast, we consider that the demand function is drawn from a parametric class of functions with unknown continuous parameters of dimension $k$. To learn the value of the parameters, we are faced with a set that has an infinite and uncountable number of elements (as opposed to the finite set in Cheung et al. (2017)). In terms of methodology, Cheung et al. (2017) apply first-order estimation (using sample average to estimate the expectation), whereas we employ Maximum Likelihood Estimation (MLE). The regret convergence rate for their model is $O(\log^{m}(mT))$ if $m$ price changes are allowed, and for our model, the best possible convergence rate (which we achieve) is $O(T^{1/(m+1)})$.

Besbes and Zeevi (2009) study the revenue management problem with fixed initial inventory by using both non-parametric and parametric approaches, thus there is no inventory decisions. For the parametric approach, they prove that the lower bound for regret of their algorithm is $\Omega(T^{1/2})$. In their $k$-unknown-parameter case (which is similar to our $k$-identifiable case), they propose an algorithm with regret $O(T^{3/2}T^{1/2})$; in their $1$-unknown-parameter case (which is similar to our well-separated case), they obtain a regret of $O(T^{1/2}(\log T)^{1/2}(\log \log T))$

1.2. Structure of this article

In the next section we formulate the joint pricing and inventory replenishment problem. In Section 3 we present...
learning algorithms, for the well-separated case and the general case as well as their regret rates. In Section 4 we conduct a numerical study and report the numerical results. We conclude in Section 5. The proofs of Theorems 1 to 4 are provided in the online Appendix.

2. Model formulation

We consider a periodic-review stochastic inventory system over a planning horizon of \(T\) periods. At the beginning of each period \(t\), the firm sets a selling price \(p_t \in \{p^1, p^2\}\) and determines a replenishment decision, or order-up-to level, \(y_t \in \mathcal{Y} = \{y^1, y^2, \ldots, y^k\}, t = 1, \ldots, T\). During period \(t\), a random number of potential customers \(D_t\) arrive, and each potential customer purchases a product with probability \(r(p_t, z)\), where \(z \in \mathcal{Z}\) is a parameter vector and \(\mathcal{Z}\) is a compact and convex set, and \(r(p, z)\) is non-increasing in \(p\). The total number of potential customers over the \(T\)-period planning horizon, \(D_1, D_2, \ldots, D_T\), are independent and identically distributed. Customers willing to purchase in a period are satisfied as much as possible by on-hand inventory, and unsatisfied demands are backlogged. Let \(b\) be the unit holding cost per period, \(h\) the unit backlog cost per period, and inventory ordering cost is normalized to zero. The inventory replenishment lead-time is assumed to be zero. The objective of the firm is to dynamically determine its pricing and inventory replenishment decisions in each period to maximize its total expected profit.

Let the potential demand distribution be denoted by

\[
P(D_t = n) = w_n, \quad n = d^l, d^l + 1, \ldots, d^u,
\]

where \(d^l\) and \(d^u\) are the lower and upper bounds, respectively, of the potential demand. For convenience we denote \(w = (w_{d^l}, \ldots, w_{d^u})\). If the firm knows the distribution of \(D_t\) as well as each customer’s response probability \(r(p, z)\) (or more specifically, the true value of \(z\)), then this is a standard dynamic joint pricing and inventory control problem that has been studied extensively in the literature. However, in our setting, the firm knows neither the distribution of \(D_t\) nor the customer response probability parameter \(z\) a priori. In addition, the firm is faced with the business constraint that prevents it from conducting extensive price exploration. Thus, the firm is subject to constraints on the number of times it can change its selling price. In such scenarios, the firm has to develop a mechanism that learns the potential demand distribution and customer response probability parameter while satisfying the constraints on the number of price changes, and exploit the extracted information to maximize its total profit.

To formulate the optimization problem, we let \(x_t\) denote the inventory level at the beginning of period \(t\) before replenishment decision, and assume \(x_1 = 0\). Clearly, given \(p_t = p\) and conditioning on \(D_t = n\), the number of customers who purchase the product in period \(t\) is a binomial random variable with parameters \(n\) and \(r(p_t, z)\). Thus, if we let \(B(n, r)\) denote a generic binomial random variable with parameters \(n\) and \(r\), then the total number of customers who purchase in period \(t\) can be written as \(B(D_t, r(p_t, z))\). Given a pricing and inventory policy \(\phi = ((p_1, y_1), (p_2, y_2), \ldots, (p_T, y_T))\) with \(y_i \geq x_i\), the total expected profit over the planning horizon is

\[
V^\phi(T) = \sum_{t=1}^{T} \left\{ p_t r(p_t, z) E[D_t] - \left\{ h E[y_t - B(D_t, r(p_t, z))]\right\}^+ + b E\left[B(D_t, r(p_t, z)) - y_t\right]^+ \right\}
\]

(1)

If the firm knows the distribution of \(D_t\) and parameter \(z\) a priori, then dynamic programming can be used to find the optimal pricing and inventory replenishment decisions that maximize problem (1). If that is the case, it is known (see, e.g., Sobel (1981)) that a myopic policy is optimal for problem (1). Therefore, to maximize the \(T\) period profit in model (1), the firm only needs to solve a single-period optimization problem that maximizes \(G(p, y, z, w)\), where

\[
G(p, y, z, w) = \frac{1}{n!} \left\{ B(p, y) - h E[y - B(p, y)]\right\}^+ - b E[B(p, y) - y]^+,
\]

and \(D\) is the generic potential random demand. Assume \(G(p, y, z, w)\) admits a unique maximizer \((p^*, y^*)\) on \(P \times \mathcal{Y}\), and \(p^* \in (p^1, p^2)\). Then the optimal policy \(\phi^*\) for the \(T\)-period problem (1) under complete information without limit on the number of price changes is to implement \((p^*, y^*)\) every period. Since the initial inventory \(x_1 = 0\), \(y^*\) can be achieved every period. As it implements the same price \(p^*\) every period, no price change is required, therefore \(\phi^*\) is also the optimal policy with limited price changes.

Let \(d_t\) and \(s_t\) denote the realized potential demand and sales in period \(t\), then the firm needs to develop an adaptive policy \(\pi\) that determines the selling price \(p_t\) and replenishment level \(y_t \geq x_t\) based on historical information \(((p_1, y_1, d_1, s_1), \ldots, (p_{t-1}, y_{t-1}, d_{t-1}, s_{t-1}))\) that satisfy the constraints on the number of price changes and carry-over inventories. We use regret to measure the performance of a policy \(\phi\), which is defined as the total profit loss of policy \(\phi\) compared with that of the optimal policy \(\phi^*\) when complete information is available and there is no constraint on the number of price changes. That is,

\[
R^\phi(T) = V^{\phi^*}(T) - V^\phi(T).
\]

For any policy \(\phi\), it holds that \(R^\phi(T) \geq 0\), and the better policy \(\phi\) performs, the smaller the regret.

We remark that in this model, both potential demand and sales data are observable. This assumption is appropriate in online setting in which the online retailers can track the total demand arrival as well as the realized purchasing decisions (hence the probability of purchasing). However, it may not hold in traditional retail stores, where the firm can only observe realized sales.

In the following section we present learning algorithms for the dynamic joint optimization of pricing and inventory
control when the firm does not have information about potential demand distribution and parameters of customer response probability \textit{a priori}, and it is subject to constraints on the number of price changes.

### 3. Learning algorithms and their regrets

To optimize profit, the firm needs to learn both the distribution of potential customer demand and the customer response to selling price (or parameter \(z\)). To estimate the true value of \(z\), we employ the MLE method.

Recall that when the selling price in period \(t\) is \(p_t\), each potential customer purchases a product with probability \(r(p_t, z)\). Let \(d_t\) be the realized customer arrivals during period \(t\) and \((u_{t1}, \ldots, u_{tdt})\) the realizations of purchasing decisions of these customers, i.e., \(u_{ti} = 1\) if customer \(i\) in period \(t\) purchases a product and \(u_{ti} = 0\) otherwise, \(1 \leq i \leq dt\). Denote

\[
d_{[t,t_2]} = (d_t, d_{t+1}, \ldots, d_{t_2}), \quad p_{[t,t_2]} = (p_t, p_{t+1}, \ldots, p_{t_2}).
\]

The likelihood function for having customer purchasing realizations

\[
u_{[t,t_2]} = (u_{ti}; t_1 \leq t \leq t_2, 1 \leq i \leq dt),
\]

between periods \(t_1\) and \(t_2\) is

\[
\mathcal{L}^{P_{[t,t_2]} \mid z}(d_{[t,t_2]}, u_{[t,t_2]}) = \prod_{t=t_1}^{t_2} \prod_{i=1}^{dt} r(p_t, z)^{u_{ti}} (1-r(p_t, z))^{1-u_{ti}}.
\]

The maximum likelihood estimator for \(z\), denoted by \(\hat{z}\), is

\[
\hat{z} = \arg \max_{z \in Z} \mathcal{L}^{P_{[t,t_2]} \mid z}(d_{[t,t_2]}, u_{[t,t_2]}).
\]

In what follows we first study a so-called well-separated case, in which the firm is either constrained by a given number of price changes, or the number of price changes is restrained to be infrequent (to be specified). Then, we consider a more general case. For each case, we provide the learning algorithm and the convergence rate of its regret.

#### 3.1. Well-separated customer response

We first consider a well-separated case, which is defined as follows. Let \(p \in [p_l, p_r]\) and \(z \in Z \subset R^l\), the probability distribution for customer purchasing decision \(u \in \{0, 1\}\) is

\[
Q^{p, z}(u) = r(p, z)^u (1-r(p, z))^{1-u}.
\]

The family of distributions \(\{Q^{p, z} : z \in Z\}\) is called well-separated if for any \(p \in [p_l, p_r]\), \(\{Q^{p, z} : z \in Z\}\) is identifiable, i.e., \(Q^{p, z_1}(\cdot) \neq Q^{p, z_2}(\cdot)\) for \(z_1 \neq z_2\). We introduce the following two assumptions for the well-separated case: For any \(p \in [p_l, p_r]\) and any \(z \in Z\), (i) there exists a constant \(c_j>0\) such that the Fisher information \(I(p, z)\), given by

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

satisfies \(I(p, z) \geq c_j\), where we use \(U\) to denote a Bernoulli random variable with distribution \(Q^{p, z}\); (ii) there exist constants \(r\) and \(\bar{r}\) such that \(0 < r \leq r(p, z) \leq \bar{r} < 1\). Examples that satisfy these assumptions include (i) \(P = [1/2, 2], Z = [1, 2]\) and logit customer response probability \(r(p, z) = e^{-\beta/(1+e^{-\beta})}\); (ii) \(P = [1/3, 1/2], Z = [3/4, 1]\) and linear customer response probability \(r(p, z) = 2/3 - zp\); and (iii) \(P = [1/2, 2], Z = [1, 2]\) and exponential customer response probability \(r(p, z) = e^{-\beta}\).

See Broder and Rusmevichientong (2012) and Chen, Jasin and Duenyas (2018) for more discussions on these conditions.

First we consider the case that the number of price changes is limited to a given number, say \(m \geq 1\). To develop a learning algorithm for this case, we divide the planning horizon \(T\) into \(m + 1\) stages, of which the \(i\)th stage consists of \(I_i = \lceil T/(m+1) \rceil\) periods, \(i = 1, \ldots, m\), and the \((m+1)\)th stage contains the last \(T-\sum_{i=1}^{m} I_i\) periods, where \([x]\) represents the smallest integer greater than or equal to \(x\). During stage \(i \geq 2\), the algorithm implements a solution that is constructed using data collected from the previous stage \(i-1\). Then the algorithm uses the realized potential demand data and realized customer purchasing data during stage \(i\) to estimate the potential demand distribution and customer response parameter \(z\), and solve a data-driven version of optimization problem for Equation (2). The resulting solution is implemented in the next stage \(i+1\).

Let \(t_i\) denote the last period of stage \(i-1\), \(i = 2, \ldots, m+2\), i.e., \(t_i = \sum_{j=1}^{i-1} I_j\) for \(i = 2, \ldots, m + 2\) and \(t_0 = 0\). Recall that \(x_t\) is the starting inventory level of period \(t, t = 1, 2, \ldots, T\).

To get the algorithm started, we need some input \(\tilde{p}_1\) and \(\tilde{y}_1\) in \(Y\) as the initial pricing and inventory decision for the first stage.

#### Algorithm I

**Step 1:** Setting pricing and replenishment decisions

For stage \(i = 1, \ldots, m+1\), set the selling price and inventory level to

\[
p_t = \tilde{p}_i, t = t_i + 1, \ldots, t_{i+1} - 1; \quad \tilde{y}_i = \max\{x_t, \tilde{y}_i\}, t = t_i + 1, \ldots, t_{i+1} - 1.
\]

**Step 2:** Estimation

For stage \(i = 1, \ldots, m\), estimate \(z\) and potential demand distribution \(w\) using the realized data in stage \(i\),

\[
p_{[t_i+1,t_{i+1}]} = (p_{t_i+1}+1 \leq t \leq t_{i+1}), \quad d_{[t_i+1,t_{i+1}]} = (d_{t_i+1}+1 \leq t \leq t_{i+1}), \quad u_{[t_i+1,t_{i+1}]} = (u_{t_i+1}+1 \leq t \leq t_{i+1}, 1 \leq d_{t_i+1} \\
\]

and \(\hat{z}_i = \arg \max_{z \in Z} \mathcal{L}^{P_{[t_i+1,t_{i+1}]} \mid z}(d_{[t_i+1,t_{i+1}]}; u_{[t_i+1,t_{i+1}]}), \quad \tilde{w}_i = (\tilde{w}_{i\alpha}, \ldots, \tilde{w}_{i\alpha}), \quad \tilde{w}_{in} = \frac{1}{t_{i+1} - t_i} \sum_{\alpha=\hat{z}_i}^{\hat{z}_{i-1}} \|d_{t_i+1} - n\|, n = d^\alpha, \ldots, d^\beta.\]

**Step 3:** Data-driven optimization

For stage \(i = 1, \ldots, m\), solve the data-driven optimization problem:

\[
\max_{(p, y) \in [p_l, p_r] \times Y} G(p, y, \hat{z}_i, \tilde{w}_i),
\]
where

\[ G(p, y, \hat{z}, \hat{w}_i) = pr(p, \hat{z}_i) \left( \sum_{n=d}^{d^i} n\hat{w}_m \right) - \sum_{n=d}^{d^i} \hat{w}_m (h\mathbb{E}[y - B(n, r(p, \hat{z}_i))]^+) + b\mathbb{E}[B(n, r(p, \hat{z}_i)) - y]^+). \]

Denote its optimal solution by \((\hat{p}_{i+1}, \hat{y}_{i+1})\), and go to Step 1.

The learning algorithm above is explained as follows. Since price cannot be changed more than \(m\) times, the planning horizon is divided into \(m + 1\) learning stages. These stages are exponentially increasing in length, and the reasoning behind it is that, as more data are collected and more accurate estimates are obtained for the potential demand distribution and customer response probability, they should be used for longer times. The algorithm contains \(m + 1\) iterations, with iteration \(i\) implementing decisions for stage \(i\) and computing the decisions for stage \(i + 1\). That is, in each iteration, the algorithm implements the pricing and inventory decision obtained from an earlier iteration, then use the data collected to estimate parameter \(\hat{z}\) and the probability distribution for potential demand. In particular, maximum likelihood method is used to estimate the parameter \(\hat{z}\), and empirical distribution is calculated for the potential demand. Then, this information is used to construct a data-driven optimization problem (4), and its optimal solutions are the pricing and inventory decisions to be implemented in the next iteration.

**Theorem 1.** There exists a constant \(c_1 > 0\) such that for any problem instance, the regret of learning algorithm I with at most \(m\) price changes is bounded by

\[ R(T) \leq c_1 T^{\frac{1}{m+1}}. \]

An important question is whether there exists a learning algorithm with \(m\) or fewer price changes that has lower regret rate than Algorithm I. The following result shows that it is not possible, at least when the times of price changes need to be determined at the start.

**Theorem 2.** There exist problem instances such that the regret for any learning algorithm that changes prices no more than \(m\) times according to a predetermined schedule is lower bounded by \(\Omega(T^{1/(m+1)})\). That is, there exists a constant \(c_2 > 0\) such that for any learning algorithm \(\phi\) that has \(m\) or fewer price changes,

\[ R^\phi(T) \geq c_2 T^{\frac{1}{m+1}}. \]

The two theorems above show that our algorithm achieves the best convergence rate for the well-separated case with a fixed number of price changes of predetermined change schedule.

It is worth pointing out that, when there is no constraint on the number of price changes, another algorithm can be developed for our problem (see Theorem 3 below) that has regret \(O(\log T)\). Therefore, the regret of the algorithm for the problem with at most \(m\) price changes increases significantly, from \(O(\log T)\) to \(O(T^{1/(m+1)})\).

In the previous analyses we consider the case that the number of price changes is constrained up to a fixed number. In practice the limit on the number of price changes may not be so stringent and more price changes may be allowed for longer planning horizons. In the following, we consider the scenario that the firm is allowed to change the price \(O(\log T)\) times and we refer to this constraint as infrequent price changes. We develop a learning algorithm for the joint pricing and inventory control problem with infrequent price changes, and show that the regret improves significantly, from \(O(T^{1/(m+1)})\) to \(O(\log T)\), due to less stringent limitations on price exploration.

For the case with infrequent price changes, we design a learning algorithm that changes price \(O(\log T)\) times. To be more specific, given input parameters \(I_0 > 0\) and \(v > 1\), the algorithm changes price for

\[ N = \left\lfloor \log_v \left( v + \frac{1}{I_0} T \right) - 2 \right\rfloor = O(\log T) \]

times. Similar to Algorithm I, the learning algorithm for this case also divides the time horizon into stages with exponentially increasing lengths, and charges the same price within each stage. However, the length of the learning stages is different, with the length of stage \(i, i = 1, 2, ..., N\), to be

\[ I_i = [I_0 v^i], \quad i = 1, 2, ..., N, \quad (5) \]

and the last stage, \(N + 1\), has \(I_{N+1} = T - \sum_{i=1}^{N} I_i\) periods. Same as for Algorithm I, define \(t_i\) to be the last period of stage \(i - 1\), i.e., \(\sum_{j=1}^{i-1} I_j = t_i, i = 2, ..., N + 2\), with \(t_1 = 0\). Thus, stage \(i\) starts in period \(t_i + 1\) and ends in period \(t_{i+1}\).

The algorithm runs in exactly the same manner as Steps 1 to 3 in Algorithm I, but with different stage length given in Equation (5), and the algorithm consists of \(O(\log T)\) iterations that increases in \(T\).

**Theorem 3.** There exists a constant \(c_3 > 0\) such that the regret of the learning algorithm with \(O(\log T)\) price changes is bounded by

\[ R(T) \leq c_3 \log T. \]

We argue that \(O(\log T)\) is the lower bound for the regret of any algorithm for our problem with at most \(O(\log T)\) (or any number of) price changes. Broder and Rusmevichientong (2012) establish such a lower bound for a special case of our problem, i.e., the dynamic pricing problem with infinite initial inventory (thus there is no inventory replenishment decision) and there is no constraint on the number of price changes, and they show that the regret for any algorithm of their problem is lower bounded by \(\Omega(\log T)\). As our problem is more general than theirs, the regret of our problem is also lower bounded by \(\Omega(\log T)\). Therefore, Theorem 3 indicates that our algorithm achieves the best possible regret rate for our problem in hand.

### 3.2. General customer response

We next consider a more general case where the parameter in the customer response probability \(r(p, z)\) is a \(k\)-
The family of distributions \( \{Q^{P^Z} : z \in \mathcal{Z} \} \) is said to belong to a general case if there exist price points \( \mathbf{p} = (p_1, \ldots, p_k) \) such that the family of distributions \( \{Q^{P^Z} : z \in \mathcal{Z} \} \) is identifiable, i.e., \( Q^{P^z}(\cdot) \neq Q^{P^z}(\cdot) \) for any \( z_1 \neq z_2 \) in \( \mathcal{Z} \). To emphasize the dependency on \( k \), in this case we shall also call the family of distributions \( k \)-identifiable, and we call \( \mathbf{p} \) the exploration prices.

We assume the following conditions for the general case: For any \( z \in \mathcal{Z} \), (i) there exists a constant \( \varphi > 0 \) such that \( \lambda_{\min}\{I(\mathbf{p}, z)\} \geq \varphi \), where \( I(\mathbf{p}, z) \) denotes the Fisher information matrix given by

\[
I(\mathbf{p}, z)_{i,j} = E_U \left[ -\frac{\partial^2}{\partial z_i \partial z_j} \log Q^{P^z}(U) \right],
\]

where \( \lambda_{\min}\{I(\mathbf{p}, z)\} \) is the smallest eigenvalue of the Fisher information matrix \( I(\mathbf{p}, z) \), and \( U \) is a vector of \( k \) independent Bernoulli random variables with joint distribution \( Q^{P^z} \); (ii) there exist constants \( \overline{r} \) and \( \underline{r} \) such that \( 0 < \underline{r} \leq r(\mathbf{p}, z) \leq \overline{r} < 1 \) for \( 1 \leq j \leq k \). These conditions are discussed in Broder and Rusmevichientong (2012) and Chen, Jasim and Duenyas (2018), and they are satisfied by a number of families of demand curves, such as (i) \( \mathcal{P} = [1/2, 2], \mathcal{Z} = [1, 2] \times [−1, 1] \) and logit customer response probability \( r(p, z) = e^{-z_1 p_1 – z_2} / (1 + e^{-z_1 p_1 – z_2}) \); (ii) \( \mathcal{P} = [1/3, 1/2], \mathcal{Z} = [2/3, 3/4] \times [3/4, 1] \) and linear customer response probability \( r(p, z) = z_1 – z_2 p \); and (iii) \( \mathcal{P} = [1/2, 1], \mathcal{Z} = [1, 2] \times [0, 1] \) and exponential customer response probability \( r(p, z) = e^{-z_1 p_1 – z_2} \). These conditions ensure that we can estimate the customer response parameter based on purchase observation at prices \( \mathbf{p} \) by the MLE method. See Besbes and Zeevi (2009) and Chen, Jasim and Duenyas (2018) for more discussions on these conditions and their implications.

To learn the \( k \)-dimensional customer response parameter \( \mathbf{z} \), we use \( k \) price changes in the learning algorithm. The algorithm, described below, divides the planning horizon \( T \) into two phases. In the first, or exploration, phase, we experiment the \( k \) exploration prices \( p_1, p_2, \ldots, p_k \), where \( \mathbf{p} = (p_1, \ldots, p_k) \) is such that \( Q^{P^z} \) is identifiable, together with an initial order-up-to level \( y \). The exploration phase consists of \( kI \) periods, where \( I = \lceil T^{1/2} / k \rceil \), and it is divided into \( k \) stages, with each stage having \( I \) periods. Thus, each of the \( k \) prices \( p_1, \ldots, p_k \) is experimented for \( I \) periods. At the end of the exploration phase, we use the collected data to compute an empirical probability mass function for the potential customer demand and to estimate the parameter \( \mathbf{z} \). These are then used to construct an empirical objective function that is optimized to find the pricing and inventory control decisions to be implemented in the second, or exploitation, phase.

**Algorithm II**

**Step 1: Exploration**

Set the pricing and order-up-to levels, for \( t = 1, \ldots, k \), to

\[
pt = \overline{p}, \quad t = (i-1)I + 1, \ldots, iI, \\
yt = y, \quad t = (i-1)I + 1, \ldots, iI.
\]

**Step 2: Parameter estimation**

Estimate the potential demand distribution and compute the maximum likelihood estimator of \( \mathbf{z} \) by realized potential customer demand data \( d^t_{[1,k]} = (d_t; t = 1, 2, \ldots, kI) \) and customer purchasing decisions \( u^t_{[1,k]} = (u_t; t = 1, 2, \ldots, kI, l = 1, 2, \ldots, d_t) \) at prices \( p^t_{[1,k]} = (p_t; t = 1, 2, \ldots, kI) \), as follows:

\[
\mathbf{z} = \arg \max_{\mathbf{z} \in \mathcal{Z}} \sum_{kI} Q^{P^z}(d^t_{[1,k]}, u^t_{[1,k]}),
\]

and \( \mathbf{w} = (\overline{w}_1, \ldots, \overline{w}_n) \), where

\[
\overline{w}_n = \frac{1}{kI} \sum_{t=1}^{kI} 1(d_t = n), \quad n = d_1, d_1 + 1, \ldots, d^k.
\]

**Step 3: Data-driven optimization and exploitation**

Solve the data-driven optimization problem:

\[
\max_{(p, y) \in [p^t_{[1,k]}] \times [y]} G(p, y, z, \mathbf{w}),
\]
where

\[ G(p, y, \hat{z}, w) = pr(p, \hat{z}) \left( \sum_{n=1}^{d} \hat{w}_n n \right) - \sum_{n=1}^{d} \hat{w}_n \left( hE[y - B(n, r(p, \hat{z}))] \right)^+ + bE[B(n, r(p, \hat{z})) - y]^+ , \]

and let the optimal solution be \((\hat{p}, \hat{y})\). For periods \(t = kl + 1, \ldots, T\), set the pricing and inventory level to

\[ p_t = \hat{p}, \quad y_t = \max \{x_t, \hat{y} \} . \]

The following theorem presents the regret rate of the profit of the learning algorithm compared with the true maximum total profit when complete demand information is available to the firm.

**Theorem 4.** For the learning algorithm of the general case, there exists a constant \(c_4 > 0\) such that for any problem instance,

\[ R(T) \leq c_4 T^4. \]

Note that in the learning algorithm described above, the pricing decisions are changed \(k\) times in order to adequately learn the \(k\) unknowns in parameter \(z\). If less than \(k\) price changes are allowed, then we will never be able to learn all the \(k\) elements in \(z\) and the regret will be linear in \(T\). If, on the other hand, \(k\) or more price changes are allowed, then our algorithm above yields a regret of \(O(T^{1/2})\), which matches the lower bound regret rate \(\Omega(T^{1/2})\) for learning algorithms of this class of problem. Indeed, we can show that even if both the pricing and inventory decisions are allowed to change in each and every period (i.e., there is no constraint on the number of changes in pricing decisions), the regret rate for any learning algorithm is lower bounded by \(\Omega(T^{1/2})\). This lower bound is established in Broder and Rusmevichientong (2012) for a dynamic pricing problem with infinite initial inventory, but it also holds in our more general setting with joint pricing and inventory replenishment decisions.

### 4. A numerical study

We consider time horizons of length \(T = 10^2, 10^3, 10^4, 10^5\), and two distributions of potential demand in each period: (i) discrete distribution taking value 0 with probability 0.1, 1 with probability 0.3, 2 with probability 0.3, 3 with probability 0.2, and 4 with probability 0.1; and (ii) truncated Poisson distribution on \(\{0, 1, 2, 3, 4, 5\}\) with parameter 2.5.

The feasible region for order-up-to level is \(Y = \{0, 1, 2, 3, 4\}\). The demand response function \(r(p, z)\) and feasible region \(P\) for selling price \(p\) are described below.

For the well-separated case, we conduct experiments when the number of price changes is constrained to 1, 2, 3, 4, or 5, or in the order of \(O(\log T)\). Starting price \(\hat{p}_1\) is randomly drawn from \(P\) according to a uniform distribution. Starting order-up-to level \(\hat{y}_1\) is randomly drawn from \(Y\), each value with probability 0.2. For each of them, we consider three customer response probabilities: (i) logic function \(r(p, z) = \exp(-zp)/(1 + \exp(-zp))\) with true value \(z = 1\), \(P = [1/2, 2]\), and \(Z = [1/5, 2]\); (ii) exponential function \(r(p, z) = \exp(-zp)\) with true value \(z = 6/5\), \(P = [1/2, 1]\), and \(Z = [1, 2]\); and (iii) linear function \(r(p, z) = 2/3 - zp\) with true value \(z = 17/20\), \(P = [1/3, 1/2]\), and \(Z = [3/4, 1]\).

For the general case, we also consider three customer response probabilities: (i) logic function \(r(p, z) = \exp(-zp - zq)/(1 + \exp(-zp - zq))\) with \(z_1 = 1\) and \(z_2 = 1 - 1, P = [1/2, 2]\), and \(Z = [1/5, 2] \times [-1, 1]\). \(p_1 = 1/2, p_2 = 3/2,\) and \(y = 3\). The number of price changes is limited to 2; (ii) exponential function \(r(p, z) = \exp(-zp - zq)\) with \(z_1 = 3/2\) and \(z_2 = 1/2, P = [1/2, 1]\), and \(Z = [1, 2] \times [0, 1]\). \(\hat{p}_1 = 1/2, \hat{p}_2 = 1,\) and \(y = 3\). The number of price changes is also limited to 2. (iii) linear function \(r(p, z) = z_1 - z_2 p\) with \(z_1 = 3/4\) and \(z_2 = 7/8, P = [1/3, 1/2]\), and \(Z = [2/3, 3/4] \times [3/4, 1]\). \(\hat{p}_1 = 1/3, \hat{p}_2 = 1/2,\) and \(y = 3\). The number of price changes is also limited to 2.

We use the following standard to evaluate the performance of the algorithm, we consider the percentage profit loss per period compared with the optimal profit of the complete information problem, which is

\[ \frac{R(T)}{G(p^*, y^*, z, w)T} \times 100\%. \]

We compute the percentage profit loss per period over 500 rounds, then report the average value in Table 1.

From Table 1, it is seen that, when \(T = 10^5\) most of percentagess of profit loss are below 10% with three exceptions, and performance improves to around 1% when \(T = 10^5\). For the well-separated case, within each column, it is seen that system performance has diminishing effect, that is, when there are initially very few price changes allowed, significant improvement can be achieved by adding one more price change, but improvement decreases when more and more price changes are allowed.

### 5. Conclusion

We consider a dynamic joint pricing and inventory control problem in which the firm has little or no prior knowledge about the distribution of potential customer demand in each period or customer response to its selling price, and in addition, the firm is subject to business constraints for extensive price exploration. We consider several scenarios and develop learning algorithms that satisfy the constraints on the number of price experiments. We obtain the regrets for these learning algorithms and show that they achieve the best possible convergence rates in the sense that, they reach the lower bound for the regrets of any algorithms for the respective classes of problems. Numerical results show that the algorithms perform very well.

In most real-world applications it is unlikely that the firm has complete information of the distribution of customer demand, hence learning is an important task for the firm’s decision-making process. In this study we consider
the scenario where the customer responses are drawn from a parametric class of distributions. It is interesting future work to explore the impact of the number of price changes on a non-parametric model.

The inventory setting considered in this article is a back- log model, i.e., demand (the customers who made pur- chases) that cannot be satisfied by on-hand inventory is backlogged. One important assumption made in this article is that all potential customers are observed, regardless of whether they purchased products or not. Although this assumption may be plausible in some applications (e.g., brick-and-mortar stores), it is conceivable that in some set- tings the retailer is unable to observe or identify customers who did not make a purchase due to a high selling price. Another direction for future research is to consider lost sales, i.e., customers that cannot be satisfied immediately are lost, and for this case there will be again two scenarios, based on whether or not the retailer can observe the lost customers, and the latter case leads to censored demand data. These all remain interesting potential directions for future research.

Acknowledgment

The authors are grateful to the Area Editor and the two referees for their helpful comments and suggestions that have helped to signifi- cantly improve the exposition of this paper.

Notes of Contributors

Boxiao Chen is an assistant professor at the College of Business Administration, University of Illinois at Chicago. Her research focuses on data-driven optimization with applications in revenue management, inventory control, and supply chain management. She holds doctoral degree in industrial and operations engineering from University of Michigan, Ann Arbor.

Xiuli Chao is a professor in the Department of Industrial and Operations Engineering at the University of Michigan, Ann Arbor. His recent research interests include stochastic modeling and analysis, inventory control, data-driven optimization, game analysis, and supply chain management. Xiuli Chao is the co-developer of the Lekin Scheduling System, and is the co-author of two books, "Operations Scheduling with Applications in Manufacturing and Services" (Irwin/ McGraw-Hill, 1998), and "Queueing Networks: Customers, Signals, and Product Form Solutions" (John Wiley & Sons, 1999). Chao shared the Erlang Prize from the Applied Probability Society of INFORMS in 1998, and he received the David F. Baker Distinguished Research Award from the Institute of Industrial and Systems Engineers (IISE) in 2005. He holds doctoral degree in operations research from Columbia University.

Funding

The research of Xiuli Chao is supported by the NSF under grant CMMI-1634676.

References

Araman, V.F. and Caldentey, R. (2009) Dynamic pricing for nonperish- able products with demand learning. Operations Research, 57(5), 1169–1188.

Aviv, Y. and Pazzal, A. (2005) A partially observed Markov decision process for dynamic pricing. Management Science, 51(9), 1400–1416.

Azoury, K.S. (1985) Bayes solution to dynamic inventory models under unknown demand distribution. Management Science, 31(9), 1150–1160.

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. (2015) On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. Management Science, 61(4), 723–739.

Bookbinder, J.H. and Lordahl, A.E. (1989) Estimation of inventory reorder level using the bootstrap statistical procedure. IIE Transactions, 21(4), 302–312.

Broder, J. (2011) Online algorithms for revenue management. Ph.D. thesis, Cornell University, Ithaca, NY.

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

Burnetas, A.N. and Smith, C.E. (2000) Adaptive ordering and pricing for perishable products. Operations Research, 48(3), 436–443.

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

Chan, L., Shen, Z., Simchi-Levi, D. and Swann, J. (2004) Coordination of pricing and inventory decisions: A survey and classification, in David Simchi-Levi, S. David Wu, & Zuo-Jun Max Shen (eds), Handbook of Quantitative Supply Chain Analysis Modeling in the E-Business Era, Chapter 9, Springer, New York, NY.

Chen, B., Chao, X. and Ahn, H.S. (2018) Dynamic pricing and inven- tory control with nonparametric demand learning. Operations Research, forthcoming.

Chen, L. and Plambeck, E. (2008) Dynamic inventory management with learning about the demand distribution and substitution probability. Manufacturing Service Operation Management, 10(2), 236–256.

Chen, Q.G., Jasin, S. and Duenyas, I. (2018) A nonparametric self- adjusting control for joint learning and optimization of multi- product pricing with finite resource capacity. Mathematics of Operations Research, forthcoming.

Chen, X. and Simchi-Levi, D. (2004) Coordinating inventory control and pricing strategies with random demand and fixed ordering cost: The finite horizon case. Operations Research, 52(6), 887–896.

Chen, X. and Simchi-Levi, D. (2012) Pricing and inventory management, in Ozalp Ozer, & Robert Phillips (eds), The Oxford Handbook of Pricing Management, Oxford University Press, pp. 784–822.

Cheung, W.H., Simchi-Levi, D. and Wang, H. (2017) Dynamic pricing and demand learning with limited price experimentation. Operations Research, 65(6), 1722–1731.

Den Boer, A. and Zwart, B. (2013) Simultaneously learning and optimiz- ing using controlled variance pricing. Management Science, 60(3), 770–783.

Ding, X., Puterman, M.L. and Bisi, A. (2002) The censored newsvendor and the optimal acquisition of information. Operations Research, 50(3), 517–527.

Elmaghraby, W. and Keskinocak, P. (2003) Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. Management Science, 49(10), 1287–1309.

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

Federgruen, A. and Heching, A. (1999) Combined pricing and inven- tory control under uncertainty. Operations Research, 47(3), 454–475.

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

Harrison, J.M., Keskin, N.B. and Zeevi, A. (2012) Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. Management Science, 58(3), 570–586.
Huh, W.T., Levi, R., Rusmevichientong, P. and Orlin, J.B. (2011) Adaptive data-driven inventory control with censored demand based on Kaplan-Meier estimator. *Operations Research*, 59(4), 929–941.

Huh, W.T. and Rusmevichientong, P. (2009) A non-parametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research*, 34(1), 103–123.

Kalish, S. (1983) Monopolist pricing with dynamic demand and production cost. *Marketing Science*, 2(2), 135–159.

Karlin, S. and Carr, C.R. (1962) Prices and optimal inventory policies, in *Studies in Applied Probability and Management Science*, Stanford University Press, Stanford, CA, pp. 159–172.

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

Lariviere, M.A. and Porteus, E.L. (1999) Stalking information: Bayesian inventory management with unobserved lost sales. *Management Science*, 45(3), 346–363.

Levi, R., Perakis, G. and Uichanco, J. (2010) The data-driven newsvendor problem: New bounds and insights. *Operations Research*, 63(6), 1294–306.

Levi, R., Roundy, R.O. and Shmoys, D.B. (2007) Provably near-optimal sampling-based policies for stochastic inventory control models. *Mathematics of Operations Research*, 32(4), 821–839.

Lovejoy, W.S. (1990) Myopic policies for some inventory models with uncertain demand distributions. *Management Science*, 36(6), 724–738.

Murray, G.R. and Silver, E.A. (1966) A Bayesian analysis of the style goods inventory problem. *Management Science*, 12(11), 785–797.

Petruzzi, N.C. and Dada, M. (2002) Dynamic pricing and inventory control with learning. *Naval Research Logistics*, 49(3), 303–325.

Scarf, H.E. (1959) Bayes solution to the statistical inventory problem. *Annals of Mathematical Statistics*, 30(2), 490–508.

Scarf, H.E. (1960) Some remarks on Bayes solutions to the inventory problem. *Naval Research Logistics Quarterly*, 7(4), 591–596.

Sobel, M. (1981) Myopic solutions of Markov decision processes and stochastic games. *Operations Research*, 29(5), 995–1009.

Subrahmanyan, S. and Shoemaker, R. (1996) Developing optimal pricing and inventory policies for retailers who face uncertain demand. *Journal of Retailing*, 72(1), 7–30.

Thowsen, G.T. (1975) A dynamic, nonstationary inventory problem for a price/quantity setting firm. *Naval Research Logistics*, 22(3), 461–76.

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.

Whitin, T.M. (1955) Inventory control and price theory. *Management Science*, 2(1), 61–68.