A heuristic policy for dynamic pricing and demand learning with limited price changes and censored demand

Citation for published version (APA):

DOI:
10.1109/SMC.2019.8914590

Document status and date:
Published: 01/10/2019

Document Version:
Accepted manuscript including changes made at the peer-review stage

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication
A heuristic policy for dynamic pricing and demand learning with limited price changes and censored demand

Jason Rhuggenaath¹, Paulo Roberto de Oliveira da Costa¹, Alp Akcay¹, Yingqian Zhang¹ and Uzay Kaymak¹

Abstract—In this work we study a dynamic pricing problem with demand censoring and limited price changes. In our problem there is a seller of a single product that aims to maximize revenue over a finite sales horizon. The seller does not know the form of the mean demand function but does have some limited knowledge. We assume that the seller has a hypothesis set of mean demand functions and that the true mean demand function is an element of this set. Furthermore, the seller faces a business constraint on the number of price changes that is allowed during the sales horizon. More specifically, the number of price changes that the seller is allowed to make is bounded above by a finite integer. We furthermore assume that the seller can only observe the sales data and thus that demand is censored. In each period the seller can replenish his inventory to a particular level. The objective of the seller is to set the best price and inventory level in each period of the sales horizon in order to maximize his profit. The profit is determined by the revenue of the sales minus holding costs and costs for lost sales (unsatisfied demand). In determining the best price and inventory level the seller faces an exploration-exploitation trade-off. The seller has to experiment with different prices and inventory levels in order to learn from historical sales data which contains information about market responses to offered prices. On the other hand, the seller also needs to exploit what it has learned and set prices and inventory levels that are optimal given the information collected so far. We propose a heuristic policy for this problem and study its performance using numerical experiments. The results are promising and indicate that the growth rate of regret of the policy is sub-linear with respect to the sales horizon.

I. INTRODUCTION

In this paper we consider a dynamic pricing and learning problem with demand censoring and limited price changes. The typical approach when tackling dynamic pricing problems with uncertain demand in revenue management, is to conduct price experimentation. By using price experimentation the seller of a product can learn the optimal price to charge for his product. Many e-commerce companies have the ability to change prices at little costs, but frequent price changes are not always desirable. For example in [1], the authors note that Groupon (a large e-commerce marketplace) does not approve price experimentation. During a selling period. Due to limited inventories, the observed sales are not equal to demand anymore. As a consequence the true underlying demand is only partially observed, that is, demand is censored.

In this paper, we consider a seller that faces demand uncertainty and has to adjust his selling price over the sales horizon in order to learn the optimal price and maximize his cumulative revenue over the sales horizon. The seller faces a business constraint on the number of price changes allowed during the selling horizon and the seller only has a limited (finite) amount of inventories on hand in each selling period. The goal of the seller is to design a pricing and inventory policy that has low regret, defined as the gap between the revenue of a clairvoyant who has full information on the demand function and the revenue achieved by a seller facing unknown demand. Demand censoring has three important implications for a seller that sets prices based on observed sales data. First, it makes demand learning harder since he does not observe true demand at a particular price. Second, the seller needs to optimize the inventory level as well, since it affects (via demand censoring) the observed sales. Third, demand censoring can lead to lost sales since some potential demand is not satisfied and this leads to lost revenues for the seller.

In this paper, we propose a heuristic policy called HPI-LPC-CD (heuristic pricing and inventory policy with limited price changes and censored demand) for the aforementioned problem of the seller. We summarize the main contributions of this paper as follows:

- We study a dynamic pricing problem with limited price changes and censored demand. In contrast with previous work, we do not assume that we can observe true demand and lost sales. Furthermore, in our setting, the lost sales part is of the objective function that the seller aims to optimize.
- We propose a pricing policy that adjust prices and inventory levels for this problem.
- We conduct numerical experiments in order to test the performance of our policy. Experimental results are promising and show that the growth rate of regret is sub-linear with respect to the sales horizon.

The remainder of this paper is organized as follows. In Section II we discuss the related literature. Section III provides a formal formulation of the problem. In Section IV we present our proposed policy. In Section V we perform experiments in order to assess the performance of our policy.
Section VI concludes our work and provides some interesting directions for further research.

II. RELATED WORK

The work in this paper is related to studies about dynamic pricing and learning with demand uncertainty, newsvendor problems, pricing with limited inventories, and pricing with limited price changes. Dynamic pricing and learning with demand uncertainty has received increasing attention in the recent years, see [2] for an extensive review. In the general dynamic pricing problem with demand learning, there is a seller that needs to decide on the optimal selling price to charge for a product. The seller, however, does not know the precise relationship between price and demand. Most of the literature (e.g. [3], [4], [5], [6], [7], [8]) assumes that there is some functional form that relates prices and demand, but that the parameters of this model are unknown, and hence, needs to be estimated from sales data. This unknown relationship between price and demand, and the objective of revenue maximization gives rise to the so-called exploration-exploitation trade-off. In order to learn the demand at various prices the seller needs to use price experimentation (exploration), but due to this price experimentation the seller is not setting the optimal price in each period and this experimentation comes at the cost of revenue maximization (exploitation).

The studies that are most relevant for the work presented in this paper are [1], [9], [10], [11]. In [1] the authors study a dynamic pricing problem with limited price changes but there is no demand censoring and no inventory decisions. The authors present a policy for this problem and characterize the regret as a function of the number of price changes. In [11], a multi-period stochastic inventory system with backlogs and demand uncertainty is considered. Although [11] do consider a setting with limited price changes, they make the assumption that potential demand (and thus the lost sales) is observed and can be backlogged. In this paper we do not make such an assumption. In [9], [10] a problem similar to the one in this paper is considered, but in [9], [10] there are no restrictions on the number of price changes. We note that demand censoring has also been studied in the setting of (repeated) newsvendor and stochastic inventory problems (e.g. [12], [13], [14]) but in these studies the demand is assumed to be stationary and there is no pricing component.

III. PROBLEM FORMULATION

In this section we give a formal description of the problem statement. The model considered here is based on the demand assumptions made in [1], but is extended to include a stochastic inventory control component similar to [12], [13], [9], [10].

Remark 1: We use the following notation in this paper. We will use \( (x)^+ \) to denote \( \max\{x, 0\} \). We use \([x]\) to denote the ceiling function applied to \( x \). We will occasionally use the symbol \( x \land y \) to denote \( \min\{x, y\} \) and the symbol \( \lor \) to denote \( \max\{x, y\} \).

We consider a monopolist seller that sells a single nonperishable product over time horizon of \( T \in \mathbb{N} \) periods. At the beginning of each period the seller can order products to adjust the on hand inventory. Furthermore, at the beginning of each period the seller has to decide on a price \( p_t \in \mathcal{P} = [p_L, p_H] \). The prices \( p_L < p_H \) are the minimum and maximum prices that are acceptable to the seller. Demand is stochastic and the seller does not know the true underlying demand model. Demand is also unobserved due to censoring, because we assume that only the sales are observed. That is, if demand exceeds the available inventory in period \( t \), then the sales equals the inventory in that period. The objective of the seller is to set prices and inventories in order to maximize revenues over the sales horizon.

A. Demand assumptions

In each period \( t \) the demand for the product is given by \( D(p) = \lambda(p) + \epsilon_t \). Here \( \lambda(p) = \mathbb{E}\{D(p)\} \) is the expected demand at price \( p \) and \( \epsilon_t, t = 1, \ldots T \) are identically and independently distributed (i.i.d) random variables with \( \mathbb{E}\{\epsilon_t\} = 0 \) and with cumulative distribution function (CDF) given by \( F(\cdot) \). We assume that \( \epsilon_t \) is bounded with known bounds \([l, u]\). The function \( \lambda(p) \) gives the expected demand at price \( p \) and we assume that \( \lambda(\cdot) \) is a non-increasing function. The function \( \lambda(\cdot) \) and the distribution of \( \epsilon_t \) is unknown to the seller and the seller has to learn these during the sales horizon.

Although the demand model is unknown to the seller, we assume that the seller does have some prior knowledge about demand. In particular, we assume that the seller has a finite hypothesis set \( \Lambda = \{\lambda_1(p), \ldots, \lambda_K(p)\} \) consisting of \( |\Lambda| = K \) mean demand functions. The demand according to hypothesis \( \lambda_k(p) \in \Lambda \) is given by \( D(p) = \lambda_k(p) + \epsilon_t \). We assume that the true mean demand function is an element of the set \( \Lambda \).

In order to distinguish between the different demand models in the hypothesis set, we make the assumption that a set of discriminative prices \( \mathcal{P}^D \) is available. A price \( p_D \in \mathcal{P}^D \) is called discriminative if \( \lambda_j(p_D) \neq \lambda_k(p_D) \) for all \( \lambda_j(\cdot), \lambda_k(\cdot) \in \Lambda \). That is, a price is called discriminative if the mean demand at that price is different for all demand functions in \( \Lambda \). The assumptions on demand that are made in this section are similar to the assumptions made in [1]. In the next subsection we discuss the assumptions made on the inventory.

B. Inventory assumptions and dynamics of the system

Let \( x_t \) and \( y_t \) denote the inventory levels at the beginning of period \( t \) before and after an inventory replenishment decision, respectively. We assume that the system is initially empty, i.e., \( x_1 = 0 \). We assume that inventory lies in a bounded interval, that is, \( y_t \in \mathcal{Y} = [y^L, y^H] \) with \( y^H \geq \lambda(0) + u \). Under these assumptions there is no demand censoring if \( y_t = y^H \).

An admissible or feasible policy is represented by a sequence of prices and order-up-to levels, \( \{(p_t, y_t), t \geq 1\} \) with \( y_t \geq x_t \), where \((p_t, y_t)\) depends only on the demand and decisions made prior to time \( t \), that is, \((p_t, y_t)\) is adapted to the filtration generated by \( \{(p_t, y_t), t \geq 1\} \) under censored demand.

Given an admissible policy \( \pi \), the following sequence of events occurs in each period \( t \):

- At the beginning of period \( t \) the seller observes the current inventory level \( x_t \).
• The seller decides to increase the inventory level to \( y_t \in \mathcal{Y} \) and decides on the price \( p_t \) that will be charged in period \( t \). Similar to previous studies (e.g. [12], [13], [11]), we assume that replenishment occurs instantly without any delay.

• The demand during period \( t \), denoted by \( d_t(p_t) \) is realized and the seller tries to satisfy as much of this demand as possible using the inventory available during the period.

• Demand that is not satisfied is lost and unobservable. More specifically, the seller only observes the sales \( \min\{d_t, y_t\} \) during period \( t \).

• At the end of period \( t \), the seller incurs a profit given by:

\[
R_t(p_t, y_t) = p_t \cdot \min\{d_t, y_t\} - b \cdot (d_t - y_t)^+ \\
- h \cdot (y_t - d_t)^+ = p_t \cdot d_t - (b + p_t) \cdot (d_t - y_t)^+ \\
- h \cdot (y_t - d_t)^+ \tag{1}
\]

Here \( b \) is a parameter that represents the costs due to lost sales and \( h \) represents the holding costs due to inventory that is left over at the end of period \( t \). Similar to previous studies (e.g. [12], [13], [11]), ordering costs are normalized to zero. Note that the profit given by Equation (1) is unobserved since it depends on the unobserved realized demand \( d_t(p_t) \).

C. Objective function under full information

The objective of the publisher is to maximize the cumulative revenue over the sales horizon of length \( T \):

\[
\max_{(p_t, y_t) \in \mathcal{P} \times \mathcal{Y}, y_t \geq x_t} \mathbb{E} \left\{ \sum_{t=1}^{T} R_t(p_t, y_t) \right\} \tag{2}
\]

Note that if \( \lambda(p) \) and the distribution of \( \epsilon_t \) was known and the seller could observe lost sales, then the optimal policy can be found by solving the following optimization problem:

\[
\max_{(p_t, y_t) \in \mathcal{P} \times \mathcal{Y}, y_t \geq x_t} \sum_{t=1}^{T} p_t \cdot \mathbb{E} \left\{ D_t(p_t) \right\} \\
- \sum_{t=1}^{T} (b + p_t) \cdot \mathbb{E} \left\{ (D_t(p_t) - y_t)^+ \right\} \\
- \sum_{t=1}^{T} h \cdot \mathbb{E} \left\{ (y_t - D_t(p_t))^+ \right\} \tag{3}
\]

However, in our setting the seller knows neither the function \( \lambda(p) \) nor the distribution of \( \epsilon_t \) and cannot observe lost-sales.

Suppose that the seller knows the function \( \lambda(p) \) and the distribution of \( \epsilon_t \). In this case it has been shown [15] that a myopic policy is optimal. We can define the single-period profit function as follows:

\[
Q(p, y) = p \cdot \mathbb{E} \left\{ D_t(p) \right\} - (b + p) \cdot \mathbb{E} \left\{ (D_t(p) - y)^+ \right\} \\
- h \cdot \mathbb{E} \left\{ (y - D_t(p))^+ \right\} \tag{4}
\]

To find the optimal pricing and inventory decision we thus need to optimize \( Q(p, y) \). The full information optimization problem (FI-OPT) that assumes that the seller knows the function \( \lambda(p) \) and the distribution of \( \epsilon_t \), can be written more compactly using (5) and (6):

\[
Z(p, \lambda(p)) = \min_{y \in \mathcal{Y}} \left\{ (b + p) \cdot \mathbb{E} \left\{ (\lambda(p) + \epsilon - y)^+ \right\} \\
- h \cdot \mathbb{E} \left\{ (y - \lambda(p) - \epsilon)^+ \right\} \right\} + (p \cdot \lambda(p)) \tag{5}
\]

\[
\max_{p \in \mathcal{P}, y \in \mathcal{Y}} Q(p, y) = \max_{p \in \mathcal{P}} \left\{ Z(p, \lambda(p)) \right\} \tag{6}
\]

D. Regret

Let the optimal solution to FI-OPT be denoted by \( (p^*, y^*) \) and the optimal single-period profit by \( Q(p^*, y^*) \). The regret of an admissible policy \( \pi \) that generates \( \{(p_t, y_t), t \geq 1\} \) can now be defined as follows

\[
\mathcal{R}(\pi, T) = T \cdot Q(p^*, y^*) - \mathbb{E} \left\{ \sum_{t=1}^{T} R_t(p_t, y_t) \right\} \tag{7}
\]

The regret measures the expected difference in revenue that arises from the fact that the seller is using policy \( \pi \) instead of the optimal policy that uses \( (p^*, y^*) \) in each period. Note that minimizing (7) is equivalent to minimizing the per period regret given by:

\[
\mathcal{R}^p(\pi, T) = Q(p^*, y^*) - \frac{1}{T} \mathbb{E} \left\{ \sum_{t=1}^{T} R_t(p_t, y_t) \right\} \tag{8}
\]

IV. PROPOSED POLICY

In this section we discuss our proposed pricing and inventory policy. First we define some preliminary notation and concepts that are needed for our policy.

A. Preliminaries

Define for each \( \lambda_k(p) \in \Lambda \) the following counterpart to FI-OPT, which we denote by \( k \)-OPT:

\[
\max_{p \in \mathcal{P}, y \in \mathcal{Y}} Q_k(p, y) = \max_{p \in \mathcal{P}} \left\{ Z(p, \lambda_k(p)) \right\} \tag{9}
\]

Here \( Q_k(p, y) \) represents the expected single-period profit (with respect to the distribution of \( \epsilon_t \)) when using the mean demand function \( \lambda_k(p) \in \Lambda \) instead of the true mean demand function \( \lambda(p) \).

Next, we define a counterpart to \( k \)-OPT that uses samples drawn from the distribution of \( \epsilon \). Suppose that we have access to \( M \) samples \( \{\epsilon_t, t = 1, \ldots, M\} \) from the distribution of \( \epsilon \). In that case we can define the sampled version of \( Q_k(p, y) \), which we denote by \( \hat{Q}^k_{SAM}(p, y) \). Using \( \hat{Q}^k_{SAM}(p, y) \) we define a sampled version of the optimization problem \( k \)-OPT, denoted by \( k \)-SAM:

\[
\max_{p \in \mathcal{P}, y \in \mathcal{Y}} \hat{Q}^k_{SAM}(p, y) = \max_{p \in \mathcal{P}} \hat{Z}(p, \lambda_k(p)) \tag{10}
\]
\[ Z(p, \lambda_k(p)) = -\min_{y \in \mathcal{Y}} \left\{ \frac{1}{M} \sum_{t=1}^{M} (b + p) \cdot (\lambda_k(p) + \hat{\epsilon} - y)^+ - h \cdot (y - \lambda_k(p) - \hat{\epsilon})^+ \right\} + (p \cdot \lambda_k(p)) \]

(11)

Note that, in general, the true demand is not observed due to possible censoring. Since the true demand is not observed due to censoring, it is useful to define a counterpart to \( k \)-SAM that is not based on samples from the distribution \( \epsilon \). Suppose that we have access to \( M \) samples \( \{\hat{y}_t, t = 1, \ldots, M\} \) that are not necessarily from the distribution \( \epsilon \). We define another sampled version of the optimization problem \( k \)-OPT, denoted by \( k \)-APPROX:

\[ \max_{p \in \mathcal{P}, y \in \mathcal{Y}} \hat{Q}_k^{APPROX}(p, y) = \max_{p \in \mathcal{P}} \hat{Z}(p, \lambda_k(p)) \]

(12)

\[ \hat{Z}(p, \lambda_k(p)) = -\min_{y \in \mathcal{Y}} \left\{ \frac{1}{M} \sum_{t=1}^{M} (b + p) \cdot (\lambda_k(p) + \hat{\eta}_t - y)^+ - h \cdot (y - \lambda_k(p) - \hat{\eta}_t)^+ \right\} + (p \cdot \lambda_k(p)) \]

(13)

B. Heuristic policy

After having defined some preliminary concepts in the previous subsection, we now proceed to present our heuristic policy. The full procedure for the heuristic policy is described in Algorithm 1. The policy takes the following parameters as input:

1) The sales horizon \( T \) and an upperbound \( m \in \mathbb{N} \) on the number of price changes that are allowed.
2) The price set \( \mathcal{P} = [p^L, p^H] \) and the order-to-levels \( \mathcal{Y} = [y^L, y^H] \).
3) An initialization parameter \( y_{\text{start}} \in \mathcal{Y} \).
4) An initialization parameter \( p_{\text{start}} \in \mathcal{P} \).
5) A fixed parameter \( 0 < v < 1 \).
6) A set of demand functions \( \Lambda = \{\lambda_1(p), \ldots, \lambda_K(p)\} \).
7) A set of discriminative prices \( \mathcal{P}^D \).
8) A constant \( C_T \) that only depends on \( T \).

The main idea of the algorithm behind our policy is to split the sales horizon into a number of phases. More specifically, the algorithm splits the sales horizon in \( m+1 \) phases. For each \( 0 \leq \ell \leq m \), a single price \( P^*_\ell \) is offered through phase \( \ell \), which starts at period \( \tau_\ell+1 \) and ends at period \( \tau_{\ell+1} \). We call phases \( 0 \) to \( m-1 \) learning phases and phase \( m \) is called the earning phase.

Except for a constant factor, the lengths of the phases are iterated-exponentially increasing. Suppose we are in phase \( \ell \) and that it runs from period \( t_1 \) to period \( t_2 \). In our policy we have that \( t_1 = \tau_\ell \) and \( t_2 = \tau_{\ell+1} = \tau_\ell + C_T [\log^{(m-\ell)}T] \) for some constant \( C_T \) that only depends on \( T \). Here \( \log^{(m)}T \) denotes \( m \) iterations of the (natural) logarithm. We define \( \log^{(m)}T = 0 \) if \( m > m^* \) where \( m^* \) is the smallest integer such that \( 0 < \log^{(m^*)}T \leq 1 \). This setup of partitioning the sales horizon is similar to the approach used in [1].

The inventory decisions are made according to the following rule:

\[ y_t = \begin{cases} y_{\text{start}}, & \text{if } t = 1 \\ y_{t-1}, & \text{if } y_{t-1} > d_{t-1}, t > 1 \\ \min\{y_{t-1} \cdot (1 + v), y^H\}, & \text{otherwise} \end{cases} \]

(14)

The main idea behind (14) is to adjust the order-up-to level upwards by a fixed factor \( v \), if a stock-out occurs. If no stock-out occurs, then we keep the order-up-to level the same as in the previous period. We note that a similar rule was also used by [12] in a stochastic inventory problem with no pricing decisions and by [9, 10] in a pricing problem with inventory decisions and unlimited price changes.

Algorithm 1 Heuristic Policy: HPI-LPC-CD

\begin{algorithm}
\begin{algorithmic}[1]
\Require \( v, p_{\text{start}}, T, y_{\text{start}}, C_T, \mathcal{P}^D, \mathcal{P}, \mathcal{Y} \).
\State Set \( \ell = 0 \).
\State Set \( \tau_\ell = 0 \).
\State Set \( P^*_\ell = p_{\text{start}} \).
\State Set \( y_1 = y_{\text{start}} \).
\For {\( \ell = 0 \) to \( m-1 \)}
\State Set \( \tau_{\ell+1} = \tau_\ell + C_T [\log^{(m-\ell)}T] \).
\State Set \( t_s = \tau_\ell + 1 \).
\State Set \( t_e = \tau_{\ell+1} \).
\If {\( t_e > t_s \)}
\State Set \( t = t_s \) to \( t_e \) do \State Compute the index \( i_\ell = \arg\min_{i \in \{1, \ldots, K\}} \left| \hat{X}_{\ell} - \lambda_i(P^*_\ell) \right| \).
\State Set \( k = i_\ell \).
\State Set \( y_\ell = \min\{d(P^*_\ell), y_1\} - \lambda_k(P^*_\ell) \).
\State Set \( (p^*, y^*) = \arg\max_{p \in \mathcal{P}, y \in \mathcal{Y}} \hat{Q}_k^{APPROX}(p, y) \) using (12) and (13).
\EndIf
\EndFor
\EndAlgorithm
\end{algorithm}

At the end of the learning phase \( \ell \) the policy computes the sample mean of the observed sales under price \( P^*_\ell \). Since \( P^*_\ell \) is a discriminative price, the seller gains information about the identity of the true demand function in this learning phase. In particular, in seller compares the observed mean sales with the mean demand at the price \( P^*_\ell \) for each demand model in the hypothesis set \( \Lambda \) (Line 15). The demand model \( k \) that is has a mean demand (at price \( P^*_\ell \)) that is the closest to the observed mean sales, is then selected to generate samples of the error-term of the demand model (Line 15 and 16). The
intuition behind this procedure is that the mean demand (at price \( P^* \)) under the true demand model should be close to the mean of the observed sales if the effect of demand censoring is not too severe. The update rule for the inventory decisions given by (14) ensures that the observed sales does not suffer too much demand censoring. The seller subsequently (Line 17) determines the optimal solution \((p^*, y^*)\) to the optimization problem given by (12) and (13). If we are not yet in the last learning phase, we select the next price to be equal to the discriminative price that is the closest to \( p^* \) (Line 19). The next value for the order-up-to level is then optimized conditional on the discriminative price that is closest to \( p^* \) (Line 20). If we are at the end of the last learning phase, then the price \( p^* \) will be used in all subsequent periods. The order-up-to level for all subsequent periods is then equal to \( y^* \).

V. NUMERICAL EXPERIMENTS

We conducted a number of numerical experiments in order to assess the performance of the policy.

A. Setup of experiments

We used the following parameter settings: \( \mathcal{P} = [0.0, 50.0], \mathcal{Y} = [0.0, 200.0], K = 8, v = 0.1 \), discriminative prices \( \mathcal{P}^D = \{5, 15, 12.5, 20, 10, 35\} \). The value of \( y_{\text{start}} \) is randomly chosen from the set \( \{10, 20, 30\} \) and the value of \( p_{\text{start}} \) is randomly chosen from the set \( \mathcal{P}^D \). Furthermore, in our experiments we set \( C_T = [10 \cdot \log T] \) in order to determine the length of the learning and earning phases.

We used the following demand functions for the hypothesis set \( \Lambda \):

\[
\exp(7.0 - 0.2p) + \epsilon \quad (15) \\
\exp(6.0 - 0.2p) + \epsilon \quad (16) \\
\exp(6.0 - 0.15p) + \epsilon \quad (17) \\
\exp(5.0 - 0.1p) + \epsilon \quad (18) \\
\exp(4.0 - 0.2p) + \epsilon \quad (19) \\
\exp(5.5 - 0.2p) + \epsilon \quad (20) \\
\exp(5.5 - 0.15p) + \epsilon \quad (21) \\
\exp(4.5 - 0.2p) + \epsilon \quad (22)
\]

In our experiments the true demand model is given by (17). This choice for the hypothesis set \( \Lambda \) models a scenario where there are multiple mean demand functions that are similar but distinct, which in turn makes it harder to the policy to identify the correct demand model.

In the experiments the errors are taken from an uniform distribution with \( \epsilon \sim \mathcal{U}(-5, 5) \). We let \( b \in \{4, 6\}, h \in \{2, 4\}, T \in \{750, 1000, 5000, 10000\}, m \in \{3, 4\} \). These choices for the parameter values of \( b \) and \( h \) model a scenario where lost sales are more costly to the seller than holding costs.

In order to better interpret the results, we report the scaled per-period regret which is given by:

\[
R^S(\pi, T) = \frac{Q(p^*, y^*) - \frac{1}{T} \mathbb{E} \left[ \sum_{t=1}^{T} R_t(p_t, y_t) \right]}{Q(p^*, y^*)} \cdot 100 \quad (23)
\]

For a particular choice of parameter settings, we average the scaled regret over 250 simulations. The results are displayed in Table I.

<table>
<thead>
<tr>
<th>( h )</th>
<th>( b )</th>
<th>( m )</th>
<th>( T )</th>
<th>Scaled regret</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
<td>750</td>
<td>7.73</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1000</td>
<td>5.79</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
<td>5000</td>
<td>0.98</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
<td>10000</td>
<td>0.26</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>750</td>
<td>7.47</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>1000</td>
<td>5.60</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>5000</td>
<td>0.94</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>10000</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>3</td>
<td>750</td>
<td>7.60</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>3</td>
<td>1000</td>
<td>5.65</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>3</td>
<td>5000</td>
<td>0.75</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>4</td>
<td>750</td>
<td>7.70</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>4</td>
<td>1000</td>
<td>5.70</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>4</td>
<td>5000</td>
<td>0.76</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>4</td>
<td>10000</td>
<td>0.32</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>750</td>
<td>9.11</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>1000</td>
<td>7.02</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5000</td>
<td>1.81</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>10000</td>
<td>0.90</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>750</td>
<td>8.83</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1000</td>
<td>6.78</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5000</td>
<td>1.76</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>10000</td>
<td>0.87</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>3</td>
<td>750</td>
<td>9.27</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>3</td>
<td>1000</td>
<td>7.08</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>3</td>
<td>5000</td>
<td>1.64</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>3</td>
<td>10000</td>
<td>0.68</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>4</td>
<td>750</td>
<td>9.16</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>4</td>
<td>1000</td>
<td>6.97</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>4</td>
<td>5000</td>
<td>1.56</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>4</td>
<td>10000</td>
<td>0.64</td>
</tr>
</tbody>
</table>

B. Results

The main insights from the experiments are as follows. The scaled regret is decreasing with respect to the sales horizon which indicates that the per-period regret is decreasing with respect to the sales horizon. The scaled regret tends to be larger for smaller sales horizons, and this makes sense, since the policy has less sales periods to determine the true underlying demand model. For smaller sales horizons the regret is as high as 10%, but as the sales horizon increases, the regret decreases relatively quickly for most of the parameter values considered. This indicates that, as the sales horizon increases, the policy is better able to make decisions that are close to optimal.

The rate at which the scaled regret decreases with the horizon appears to be related to (the difficulty of) the problem instance. In particular, when \( h = b = 4 \) the scaled regret decreases at a slower rate compared to the case when \( h = 2 \) and \( b = 4 \). One possible explanation is that, when \( h = 2 \) and \( b = 4 \), the
policy recognizes that lost sales are more costly and it more obvious for the policy that it needs to learn the right values for the inventory decisions quickly. However, when $h = b = 4$, their is less incentive for the policy to learn the right values for inventory decisions quickly.

Figure 1 plots the values of $T \cdot RS(\pi, T)$ against the sales horizon $T$. This figure is useful in order to visualize of how the regret scales with the sales horizon $T$. The patterns in Figure 1 illustrate that the regret generally increases slowly with the horizon, and in some cases, also decreases. These patterns confirm what we already saw in Table I and shows that, as the sales horizon increases, the per-period regret tends to decrease.

Overall, the results suggest that the growth rate of regret is sub-linear in the sales horizon $T$. The results are promising and indicate that the policy is able to learn the true demand model and set the right values for the price and inventory.

![Fig. 1: Values of $T \cdot RS(\pi, T)$ against the sales horizon $T$.](image)

VI. CONCLUSIONS

In this paper we studied a dynamic pricing problem with limited price changes and censored demand. In contrast with previous work, we did not assume that we can observe true demand and lost sales. Furthermore, in our setting, the lost sales is part of the objective function that the seller aims to optimize. We proposed a heuristic pricing policy that adjust prices and inventory levels for this problem. Using numerical experiments we tested the performance of our policy. Experimental results are promising and suggest that the growth rate of regret is sub-linear with respect to the sales horizon.

Future work could be directed towards deriving analytical results related to the dynamic pricing problem studied in this paper. In particular, it would be interesting to study upper and lower bounds on the growth rate of regret and how these bounds depend on the number of price changes and the sales horizon.

REFERENCES