POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

A Multi-Armed Bandit Approach to Dynamic Pricing

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Author: Gianmarco Genalti

Advisor: Prof. Nicola Gatti

Co-advisors: Marco Mussi, Alessandro Nuara, Ph.D.

Academic year: 2020-2021

1. Introduction

Dynamic Pricing refers to a family of techniques used to learn the optimal price of a product or service in a real-time fashion. Dynamic pricing is receiving significant attention from both the industry and scientific community due to the economic impact that such algorithms have on business and the many theoretical challenges they have to face to sound enough to be put into production. The widespread of e-commerces and marketplaces resulted in a synergistic speed up in data collection and availability and *data-driven* dynamic pricing algorithms development. It is not excessive to consider these two categories of retailers as main users for a dynamic pricing strategy.

1.1. Goal

Our goal is to design a framework that enables companies to perform data-driven dynamic pricing on their products while guaranteeing the scientific soundness of the deployed algorithm. Moreover, we add a new layer to the typical dynamic pricing algorithm's workflow by proposing a novel data-driven volume discounts policy that works with the pricing. We propose a novel *Online Learning* methodology, aiming at reduce the gap between industrial needs and the scientific state-of-the-art. While most of the works on dynamic pricing do not present a real-world experimental campaign as a validation of their proposal, we had the possibility to test our algorithm for more than 4 months by pricing products for a total turnover of 160 KEuros. The test was conducted on an Italian e-commerce selling consumer products.

1.2. Context

In this work, we focus on a monopolistic pricing problem on e-commerce in the presence of volume discounts and seasonality, where the objective function is a convex combination between margins and turnover and only transaction data are available.

Industry International reports quantify that more than 14 trillion USD per year, since 2030, will be unlocked thanks to automation of industrial and business processes. Automation of pricing tasks, in particular, is estimated to unlock value for about 0.5 trillion of USD per year worldwide. The nature of the objective function is of paramount importance for industrial applications, because it allows companies to decide what are their goals and market aggression policy. In industrial context, scarcity of data

in one of the most common problems to deal with, the online nature of our proposal naturally deals with this. Moreover, the joint learning of customers' demand and seasonality enforce it against outliers as well as the scarcity of data mentioned above.

Research Research on this topic is split among different scientific communities: from classic economics to *online learning*. Each one has its own point of view and, consequently, its focus. While econometry approaches tend to investigate the nature of the market assumptions and how they affect pricing, statisticians will be interested in the mathematical modelling of the demand curve. Management science and operational research come across these two fields and try to unify them. In contrast, online learning and reinforcement learning communities will try to investigate the theoretical performances of an agent which sets prices following a given policy.

1.3. Original Contributions

We decompose sales time series into two components: a base demand and the price elasticity (price-induced volumes). Base demand is composed mainly of trend and seasonality, while price elasticity is the realization of the demand curve, namely the relationship between price and units sold. Motivated by goods different from luxury, Veblen, and Giffen, we assume that the demand curve is monotonically decreasing in the price. We force such an assumption in the learning algorithm. To do so, we design a novel Bayesian regression algorithm that forces a subset of features to be monotonic. Next, the exploitation-exploration dilemma is faced by using a Multi-Armed Bandit approach, choosing Thompson Sampling as exploration strategy to exploit the uncertainty provided by the Bayesian modelling. Our algorithm also computes volume discounts, adapting such discounts to users need, given the buyback probability. We name our algorithm PSV-B (Pricing with Seasonality and Volume discounts Bandit algorithm). Finally, a real-world A/B test has been performed. Historical data were initially fed to the algorithm to make it learn the base demand, then the algorithm learned during the 4 months the relationship between pricing and volumes. At the end, our algorithm PSV-B provided a performance that is better than the performance of B configuration for about 55%.

2. Methodologies

In this chapter, we introduce the main theoretical tools used in this work, in particular: Linear Basis Function Models and their extension to a Bayesian framework with Bayesian Linear Regression (BLR) [1] and the Multi-Armed Bandit (MAB) problem [3] (in order to remain in a Bayesian setting we propose Thompson Sampling as a framework to deal with the exploration-exploitation dilemma).

2.1. Bayesian Linear Regression

The goal is to obtain a linear relationship between the response variable $y \in \mathbb{R}$ and the features $\boldsymbol{x} \in \mathbb{R}^D$, to better generalize this relationship we introduce (nonlinear) basis functions $\Phi(\mathbf{x}) = (\Phi_0(\mathbf{x}), \dots, \Phi_{M-1}(\mathbf{x}))$, we can now compactly write:

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \Phi(\mathbf{x}) \tag{1}$$

In a Bayesian framework, BLR is obtained by putting a prior probability distribution over the weights of the regression. For our purposes we consider a *self-conjugated* distribution $\mathcal{D}$ having $\boldsymbol{m}_0 \in \mathbb{R}^M$ as mean and $S_0 \in \mathbb{R}^{M \times M}$ and variance:

$$p(\mathbf{w}) = \mathcal{D}(\mathbf{m}_0, \mathbf{S}_0) \tag{2}$$

If N samples are available, posterior distribution can be obtained as a product of the prior and the likelihood, with resulting mean and variance of $\boldsymbol{m}_N \in \mathbb{R}^M$ and $S_N \in \mathbb{R}^{M \times M}$ respectively:

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{D}(\mathbf{m}_N, \mathbf{S}_N) \tag{3}$$

While choosing a *Gaussian* prior allow us to get a closed form solution of the posterior, other types of prior may require a sampling or variational inference approach for the posterior estimation.

Online Learning and BLR If data points arrive sequentially, then the posterior distribution at any stage acts as the prior distribution for the subsequent data point. Thanks to this property, we can easily use BLR in an *Online Learning* setting, where data arrive from time to time and we do not need to train the whole model every time.

Isotonic BLR We put ourselves in the scenario in which the relationship between target variable and input features is known to be monotonic due to the physical process involved. This setting is of particular interest for dynamic pricing, since demand curve can be assumed to be decreasing in price for non-luxury products. In a BLR setting, we use a particular basis function expansion called *Bernstein Polynomial expansion*. The k-th Bernstein Polynomial basis function of order M is defined as

$$\psi_k(x, M) = \binom{M}{k} x^k (1-x)^{M-k}, \quad x \in [0,1] \tag{4}$$

Using a particular linear transformation $\boldsymbol{\theta} = A\boldsymbol{w}$ [2] on the weights of the regression, the formulation results in:

$$f(x) = \boldsymbol{\Psi} A^{-1} \boldsymbol{\theta} \tag{5}$$

Now, we can obtain a monotonically increasing function by imposing $\theta_k \geq 0$ for all $k = 1, \ldots, M$. In order to obtain a downward monotonic regression model is sufficient to flip the basis functions using $1 - \Psi A^{-1}$ instead of ΨA^{-1}. Finally, if the probability distributions over the transformed coefficients $\boldsymbol{\theta}$ have positive support we are intrinsically forcing a positive value on them. These distributions, jointly on decreasing basis functions allow us to obtain an isotonic, downward regression.

3. Problem Formulation

We study the scenario in which a monopolistic e-commerce website sells a non-perishable product with unlimited availability and the demand function is monotonically decreasing in the price, possibly nonstationary.

3.1. Pricing Formulation

At every time t, with an arbitrary granularity, we are faced with the choice of a, potentially different, price p_t from a finite set $\mathcal{P}$. The actual average number of sales (a.k.a. volumes) at time t when choosing price p_t is denoted with $v_t(p_t)$. In particular, we assume that the volumes depend on both price and time due to, *e.g.*, seasonality and market trend, and we denote the volumes curve function with $\mathcal{V}(t, p_t)$, where $\mathcal{V} : \mathcal{T} \times \mathcal{P} \to \mathbb{R}^+$. At every time t, we have $v_t := \mathcal{V}(t, p_t)$. Finally, every unit sold, the agent gains a margin $m_t := p_t - c$, where $c \in \mathbb{R}^+$ is the cost of the product. The cost of a product is assumed to be constant: in this corner scenario, customers' reaction to price and to net margin are assumed to be equal, allowing us to study price-elasticity effect directly on margins. The objective function to maximize is defined as a convex combination with parameter $\lambda \in [0,1]$ between *turnover* and *operating cash flow margin*. Formally, the maximization problem is as follows:

$$p_t^* = \underset{p_t \in \mathcal{P}}{\operatorname{argmax}}\, f(p_t), \tag{6}$$

where:

$$f(p_t) = \lambda \frac{p_t \; v_t(p_t)}{\max_{p_t \in \mathcal{P}} \{p_t \; v_t(p_t)\}} + \\ + (1-\lambda) \frac{m_t \; v_t(p_t)}{\max_{p_t \in \mathcal{P}} \{m_t \; v_t(p_t)\}} \tag{7}$$

In Eq. (7) we are balancing turnover and operating cash flow margin at time t when choosing price p_t. $\max_{p_t}\{p_t \; v_t(p_t)\}$ and $\max_{p_t}\{m_t \; v_t(p_t)\}$ are the maximum achievable values for the two measures respectively.
In real-world scenarios, functions $\mathcal{V}(t, p_t)$ and, *a fortiori*, $v_t(p_t)$ are not *a priori* known and need to be estimated online. Thus, our problem can naturally formulated as an online learning problem where the goal is to properly balance the acquisition of information on the stochastic functions, while minimizing the cumulative regret. Such a problem is also commonly known as exploration-exploitation dilemma. In our case, the arms are the possible values of price $p_t \in \mathcal{P}$, while $\mathcal{V}(\cdot, \cdot)$ and $v_t(\cdot)$ is a stochastic function that we need to estimate during the time horizon T.

3.2. Volume Discounts Formulation

Customers can purchase a single time from the shop or return many times, generating more orders. Retailers' interest is to match the volume needs of loyal customers trying to avoid their abandon. A common way to deal with these issues is to provide volume discounts, *i.e.* providing different prices depending on the number of units bought by the customer. Assuming the customer's need for units of the same product is fixed and is equal to N, the goal is to sell him

as much as possible before he drops out of the shop. The probability of repurchase γ can be estimated in multiple ways via analysis of the transaction data.
Consider a vector of η volume thresholds $\boldsymbol{\omega} = [\omega_1, \omega_2, \dots, \omega_\eta]$, with $\omega_i > \omega_h, \forall i > h$ and $\omega_1 = 1$. The price of the product is a piecewise constant function of the volume, which assigns the same price to all volumes between two consecutive volume thresholds. Let $p_t^{(i)}$ denote the price associated with the volumes between ω_i (included) and ω_{i+1} (excluded).
The goal is to define the discount δ_i that we can apply to the price for a unit volume ($\bar{p} = c + \bar{m}$) in order to get the price for the i-th volume range. To avoid negative margins, we apply the discount directly to the margin: $p_t^{(i)} = c + (1 - \delta_i)\bar{m}$. The discount δ_i should guarantee, for a customer who needs N product units and has a buyback probability γ, that the expected margin with multiple-unit orders is no lower than the one obtainable with N single-unit orders.

4. Algorithm

The goal of this algorithm is to propose a pricing schedule for a given product. A pricing schedule consists of a sequence of prices coupled with volume thresholds, namely a minimum number of units to be purchased to access the corresponding price. At each time t, the algorithm receives transaction data collected in $t-1$ and promptly computes a new pricing schedule modifying the current one.

4.1. Pricing without Volume Discounts

The pricing algorithm is based on a demand curve model fitted using a BLR.
The input space is denoted with $\mathcal{T} \times \mathcal{P}$, representing the possible combinations of time and price. Instead, the output space is denoted with $\mathcal{V}$, representing the volume. Furthermore, we introduce two features spaces $\mathcal{U}$ and $\mathcal{D}$, corresponding to seasonality&trend and price. We define $\mathcal{J}$ and $\mathcal{K}$ as the sets containing the indices of time and price features, respectively.
In particular, we introduce the function $\chi : \mathcal{T} \to \mathcal{U} \subset \mathbb{R}^{|\mathcal{J}|}$ mapping a time t into its seasonality&trend features and the function $\xi : \mathcal{P} \to \mathcal{D} \subset \mathbb{R}^{|\mathcal{K}|}$ mapping a price p_t into its basis function expansion. While $\mathcal{J}$ is represented in polar coordinates in order to ensure consistent behavior between the periods, $\mathcal{K}$ is composed by transformations that are actually monotonically decreasing in order to model the inverse relation binding price and volumes.
We force features' weights to be positive by choosing *Lognormal* priors for features in $\mathcal{K}$, while we use standard Gaussian priors for the ones in $\mathcal{J}$.

4.2. Exploration Strategy

We resort to Thompson Sampling (TS). By construction, a Bayesian model provides a probability distribution of the posteriors on the weights. TS randomly generates samples from the posterior distribution of the weights of BLR, retrieving in this way a realization of the posterior binding features from time and price to the volumes' curve. Now, given time h, fixing related features vector, we can evaluate volumes with respect to only to price values. Consider a MAB approach in which we select the best arm over a finite set of possible prices (representing the arms) $\mathcal{P}$: the selected arm is the one maximizing Eq. 7 where v_t is the TS realization.

4.3. Pricing with Volume Discounts

Let η be the desired number of volumes thresholds to propose along with as many different prices. Let β_z, with $z \in \mathbb{N}$, be the probability that a basket containing the product contains it in z units. We define the average volume inside a threshold as $\bar{V}_k$ and the total average volume of the product is $\bar{V}$. If m_t^* is the optimal margin to apply at time t, obtainable using the previously introduced method, then we can compute an optimal pricing schedule involving volume discounts. First we define the optimal margin for threshold ω_1 as

$$\bar{m}_1 = \frac{m_t^* \bar{V}}{\sum_{k=1}^{\eta} (1 - \delta_k) \bar{V}_k}, \tag{8}$$

where:

$$\delta_k = 1 - \frac{1 - \gamma^N}{\bar{V}_k \left(1 - \gamma^{\left\lceil \frac{N}{\bar{V}_k} \right\rceil}\right)}, \tag{9}$$

then the margins $\bar{m}_1, \bar{m}_2, \dots, \bar{m}_\eta$ for the different volume thresholds are determined by $\bar{m}_k = \bar{m}_1(1 - \delta_k)$, for $k = 1, \dots, \eta$. In this formulation, the expected revenue using a volume dis-

counts policy cannot be lower than the one without, but with the advantage to mitigate customers' dropout selling them more units before they leave.

5. Experimental Evaluation

The algorithm have been evaluated in both simulated and real-world scenarios.

5.1. Evaluation in Synthetic Environment

The dynamic pricing algorithm proposed is compared to the non-shape-constrained version of the algorithm, which consider *Normal* priors on the weights.

Robustness to Noise and Outliers One of the main advantages that we expect to gain using a shape-constrained algorithm is the better robustness to stochastic perturbations of the environment and outlier samples. We test the two algorithms in the same scenario, using the same random seed and the same demand curve generating function: in the simulation we assume the base demand (seasonality and trend) to be fixed and known and we focus on the hidden demand curve learning, defined as $f(x) = 2e^{-(x+1.2)^{\frac{5}{2}}}$. Setting parameter $\lambda = 0$, the reward function results in $(x-c)f(x)$. We perform a simulation for each combination of noise value and % of outliers among the generated data. The grid is generated using noise values in $\{0.001, 0.005, 0.01\}$ and outlier % in $\{0, 10, 20\}$. In Table 1, we report the % improvement in total regret moving from the free shape model to the shape-constrained one: shape-constrained model manage to always outperform the other one.

Noise	**Outlier Proportion**		
	0	**10%**	**20%**
0.001	-25%	-23%	-23%
0.005	-52%	-15%	-14%
0.01	-55%	-4%	-1.8%

Table 1: Total Regret's improvements from free shape to shape-constrained model.

Robustness to Nonstationarity In real-world scenarios, another great issue to deal with is the intrinsic nonstationarity of the customers' demand. Demand curve shape may change over time, usually in an abrupt way. We assume four possible demand functions $f_1(x)$, $f_2(x)$, $f_3(x)$ and $f_4(x)$ and we abruptly change the one generating the data during a simulation. We approach this task using a sliding window approach and tuning the window size. We consider the scenario in which the number of abrupt changes may vary in $1, 2, 3$ and the size of the sliding window in $\{20, 30, 40\}$. In Table 2 we report the % improvement in total regret moving from the free shape model to the shape-constrained one.

Window Size	**Demand Curve Changes**		
	1	**2**	**3**
20	-17%	-33%	-36%
30	-3%	-24%	-47%
40	-8%	-26%	-49%

Table 2: Total Regret's improvements from free shape to shape-constrained model.

The shape-constrained model manage to always outperform the other one, especially in the very less stationary environments.

5.2. Evaluation in Real-World Environment

We deployed our algorithm on an italian e-commerce selling consumer goods (non-Giffen). In this real-world experiment, we optimize the prices in presence of volume discounts, computing a full pricing schedule. To evaluate the algorithms performance we perform an online A/B test campaign.

Experimental Setting The experimental campaign is conducted in one of the main category of the e-commerce, with a test set (A) composed by $N_t = 295$ products and a control set (B) composed by $N_c = 33$ products of the same category with the same characteristics. The test is about products with a turnover of 300 KEuros and a net margin of 83 KEuros. E-commerce specialists asked us to impose $\eta = 3$ volume's thresholds to each product of the test set, and optimize the corresponding proposed discounts. The test lasted for 17 weeks, from 16 June 2021 to 17 October 2021, updating the prices each 7 days. There were no factors that can influence the performance of the test set (A) w.r.t. con-

trol set (B) and vice-versa (*i.e.* variations in advertising expenditures).

Performance Metric The business goal is to maximize the net margin (*i.e.* $\lambda = 0$ in Equation (7)).

Experiment Results The goods priced by PSV-B performed an improvement in the performance metric of +55% w.r.t the control set of goods. Looking at the performances on a product-wise level, we report in Figure 1 the sorted % of improvement on the performance metric w.r.t. to the period of 2021 preceding the test, for each single product.
In the test set, 138 products over 295 ($\sim$ 47%) improved their average performance with respect to the previous period of 2021, while in the control set only 8 products over 33 ($\sim$ 25%) did so.

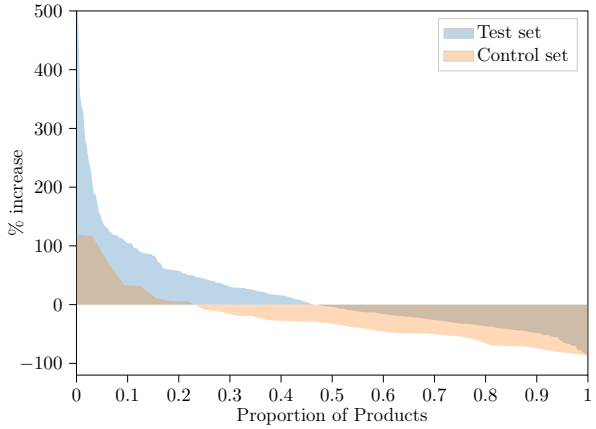

Figure 1: Margin Improvement over products.

Effect of Volume Discounts The goal of the volume discounts algorithm is to modify the probability distribution of the units count of the same product in a basket. The goal is to move mass from $\bar{\beta}_1$ to $\bar{\beta}_2$ or $\bar{\beta}_3$. In Table 3 are reported the variations of the three $\bar{\beta}_k$ of 4 relevant products: during the test we achieved an improvement of $\bar{\beta}_2$ and $\bar{\beta}_3$ at the expense of $\bar{\beta}_1$. Volume discounts not only modify cart distribution, but also increase the number of units purchased per time.

Product	$\Delta\bar{\beta}_1$	$\Delta\bar{\beta}_2$	$\Delta\bar{\beta}_3$
1	-32%	+10%	+22%
2	-26%	+25%	+1%
3	-15%	+4%	+11%
4	-5%	+1%	+4%
Mean	-19.5%	+10%	+9.5%

Table 3: Variations of $\bar{\beta}_k$ after test.

6. Conclusions

We introduced a novel dynamic pricing algorithm to deal with typical real-world scenarios. This solution is able to produce a pricing schedule accounting for seasonality and integrating a data-driven volume discounts policy. We validated design choices in a synthetic environment, then a real-world experimental campaign has been conducted to assess the added economic value that our algorithm can provide. We performed an online A/B test on an italian e-commerce lasted for more than 4 months. The algorithm improved by 55% the net cash flow margin w.r.t. the set B. Volume discounts policy positively impacted customers' shopping behaviors. We reported the example of 4 significant products having their average units per shopping cart increased by +33% with respect to the past.

References

[1] Christopher M Bishop. Pattern recognition. *Machine learning*, 128(9), 2006.

[2] S McKay Curtis and Sujit K Ghosh. A variable selection approach to monotonic regression with bernstein polynomials. *Journal of Applied Statistics*, 38(5):961–976, 2011.

[3] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.