Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Nonparametric Pricing Analytics with Customer Covariates

Ningyuan Chen , Guillermo Gallego

informs
http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 69, No. 3, May–June 2021, pp. 974–984
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Nonparametric Pricing Analytics with Customer Covariates

Ningyuan Chen,[a] Guillermo Gallego[b]

[a] Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada; [b] Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

Contact: ningyuan.chen@utoronto.ca, https://orcid.org/0000-0002-3948-1011 (NC); ggallego@ust.hk, https://orcid.org/0000-0002-9664-3750 (GG)

Received: May 3, 2018
Revised: December 21, 2019
Accepted: February 14, 2020
Published Online in Articles in Advance: February 15, 2021

Area of review: Stochastic Models
OR/MS Subject Classifications: statistics: nonparametric; inventory/production: policies: pricing

https://doi.org/10.1287/opre.2020.2016

Abstract. Personalized pricing analytics is becoming an essential tool in retailing. Upon observing the personalized information of each arriving customer, the firm needs to set a price accordingly based on the covariates, such as income, education background, and past purchasing history, to extract more revenue. For new entrants of the business, the lack of historical data may severely limit the power and profitability of personalized pricing. We propose a nonparametric pricing policy to simultaneously learn the preference of customers based on the covariates and maximize the expected revenue over a finite horizon. The policy does not depend on any prior assumptions on how the personalized information affects consumers' preferences (such as linear models). It adaptively splits the covariate space into smaller bins (hyper-rectangles) and clusters customers based on their covariates and preferences, offering similar prices for customers who belong to the same cluster trading off granularity and accuracy. We show that the algorithm achieves a regret of order $O(\log(T)^2 T^{(2+d)/(4+d)})$, where T is the length of the horizon and d is the dimension of the covariate. It improves the current regret in the literature (Slivkins 2014) under mild technical conditions in the pricing context (smoothness and local concavity). We also prove that no policy can achieve a regret less than $O(T^{(2+d)/(4+d)})$ for a particular instance and, thus, demonstrate the near-optimality of the proposed policy.

Funding: N. Chen received financial support from the Hong Kong Research Grant Council [Early Career Award 26201617] and the startup funds of the University of Toronto. G. Gallego received financial support from the Hong Kong Research Grant Council [Project 16204117 and 16211118].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2020.2016.

Keywords: multi-armed bandit • dynamic pricing • online learning • regret analysis • contextual information

1. Introduction

Personalized pricing refers to the practice that a firm charges customers different prices for the same product, depending on customers' information, such as education backgrounds and zip codes. It is increasingly popular in online retailing, as sellers can acquire/infer the personalized information from customers' account profiles or browsing histories (cookies). The demand (purchasing probability) of each customer depends not only on the price, but also on the personalized information. The firm observes the information of each arriving customer and sets a personalized price accordingly. We are interested in finding pricing policies of the firm that maximize the long-run revenue.

Personalized pricing presents several challenges to the firm. First, for new entrants to the online business, the demand function and how it depends on the personalized information is generally unknown. Thus, the optimal pricing cannot be obtained by directly solving an optimization problem. The firm may experiment with different prices to learn the personalized demand function and then sets optimal prices according to the estimation. There is usually a finite horizon that forces a trade-off between gathering more information (learning/exploration) and making sound decisions (earning/exploitation). This problem, sometimes referred to as the learning/earning dilemma, has attracted the attention of many researchers.

A second challenge is the presence of personalized information, or *covariates*. On one hand, the covariate of each customer provides extra information for the firm to predict the personalized demand more accurately. On the other hand, the demand is peculiar to each instance and changes over time, which adds significant complexity to the learning problem described above. In particular, learning a market-wise demand function is not sufficient, and the firm has to learn the personalized demand by experimenting with prices for customers of similar profiles.

A third challenge is the selection of a predictive model to estimate the demand. Consider a toy example: The demand function only depends on the address of each customer and nothing else. The firm may postulate a linear model

$$\text{demand} = a - b \times \text{price} + \boldsymbol{c} \cdot (\text{latitude, longitude}),$$

and uses historical sales data to learn the parameters a, b, and $\boldsymbol{c}$ and maximize revenue according to the estimation. However, whether the model is specified correctly plays an important role in the performance of the pricing policy. In the particular example, linearity in the location (latitude and longitude) implies that the customers along the straight line that is orthogonal to $\boldsymbol{c}$ have the same demand. This hardly reflects the reality, as customers from the same neighborhood tend to have similar shopping patterns, and neighborhoods are usually clustered geographically. By postulating a parametric (linear) model, the firm faces the risk of misspecification and not learning what is supposed to be learned. A *nonparametric* model is more appropriate in this setting.

In this paper, we study the pricing policy of a firm which tries to maximize the unknown expected revenue $f(\boldsymbol{x}, p) \triangleq pd(\boldsymbol{x}, p)$, where p is the price, and $d(\boldsymbol{x}, p) \in [\mathbf{0}, \mathbf{1}]$ is the personalized demand function for a customer of covariate $\boldsymbol{x}$. We assume both quantities are normalized, so $\boldsymbol{x} \in [\mathbf{0}, \mathbf{1})^d$ and $p \in [0, 1]$. In period t, the firm observes an arriving customer with a random covariate $\boldsymbol{X}_t$ and sets a personalized price p_t. The earned revenue is p_t multiplied by a Bernoulli random variable with success rate $d(\boldsymbol{X}_t, p_t)$, representing the event of a purchase. The expected revenue is thus $f(\boldsymbol{X}_t, p_t)$.

We propose a nonparametric learning policy for the firm. That is, the policy does not depend on specific forms of $f(\boldsymbol{x}, p)$ and only assumes general structures, such as continuity and smoothness. The policy achieves near-optimal performance compared with a clairvoyant who knows $f(\boldsymbol{x}, p)$ and sets $p^*(\boldsymbol{x}) = \arg\max_p f(\boldsymbol{x}, p)$ for a customer of covariate $\boldsymbol{x}$. More precisely, the expected difference in total revenues between the proposed policy and the clairvoyant policy, which is referred to as the *regret* in the literature, grows at $O(\log(T)^2 T^{(2+d)/(4+d)})$ as $T \to \infty$. The rate is sublinear in T, implying that when the length of horizon tends to infinity, the average regret incurred per period becomes negligible. Moreover, we prove that no pricing policies can achieve a lower regret than $O(T^{(2+d)/(4+d)})$ for a reasonable class of unknown objective functions f. Therefore, we successfully work out the learning/earning dilemma with covariates.

The main contribution of the paper is the design of a near-optimal nonparametric learning policy for the personalized pricing problem. Nonparametric learning policies are introduced in more general settings by Rigollet and Zeevi (2010), Perchet and Rigollet (2013), and Slivkins (2014). The formulation in this paper is originally introduced in Slivkins (2014), and our policy builds upon the idea of adaptive binning proposed in Perchet and Rigollet (2013).[1] Motivated by the application of personalized pricing, we assume that the expected revenue $f(\boldsymbol{x}, p)$ is smooth and locally concave in the charged price p. This deviates from the Lipschitz continuous condition in Slivkins (2014) and can be viewed as a special "margin condition" in Rigollet and Zeevi (2010) and Perchet and Rigollet (2013). By utilizing this condition, we are able to show that our policy is near-optimal and achieves improved regret over merely continuous objective functions ($T^{(\mathbf{2}+d)/(\mathbf{4}+d)}$ versus $T^{(2+d)/(3+d)}$ in Slivkins 2014).

1.1. Literature Review

This paper is motivated by the recent literature that analyzes a firm's pricing problem when the demand function is unknown (e.g., Araman and Caldentey 2009, Besbes and Zeevi 2009, Farias and Van Roy 2010, Broder and Rusmevichientong 2012, den Boer and Zwart 2014, Keskin and Zeevi 2014, and Cheung et al. 2017). den Boer (2015) provides a comprehensive survey for this area. Because the firm does not know the optimal price, it has to experiment with different (suboptimal) prices and update its belief about the underlying demand function. Therefore, the firm has to balance the exploration/exploitation trade-off, which is usually referred to as the learning-and-earning problem in this line of literature. Our paper considers the pricing problem with personalized information, and it does not consider the finite-inventory setting, as in some of the papers mentioned above.

More recently, several papers investigate the pricing problem with unknown demand in the presence of covariates (Javanmard and Nazerzadeh 2016, Nambiar et al. 2016, Qiang and Bayati 2016, Ban and Keskin 2021, Cohen et al. 2020). The existing literature has adopted a parametric approach: For example, the actual demand can be expressed in a linear form $\alpha^T \boldsymbol{x} + \boldsymbol{\beta}^T \boldsymbol{x} p + \boldsymbol{\epsilon}$ (Qiang and Bayati 2016, Ban and Keskin 2021), where $\boldsymbol{x}$ is the feature vector of a customer, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectorized coefficients, and $\boldsymbol{\epsilon}$ is the random noise. Because of the parametric form, a key ingredient in the design of the algorithms in this line of literature is to plug in an estimator for the unknown parameters (α and β), in addition to some form of forced exploration. In contrast, we focus on a setting where the demand function cannot be parametrized. Thus, the firm cannot count on accurately estimating the function globally by estimating a few parameters. Instead, a localized optimal decision has to be made based on past covariates generated in the neighborhood. It highlights the different philosophies when designing algorithms for parametric/nonparametric learning problems with covariates. As a result, the best achievable regret deteriorates from $O(\sqrt{T})$ or $O(\log T)$ (parametric) to $O(T^{(2+d)/(4+d)})$ (nonparametric).

The dependence of the optimal rate of regret on the problem dimension d has been observed before. For example, Cohen et al. (2020) find a multidimensional binary search algorithm for feature-based dynamic pricing, which has regret $O(d^2 \log(T/d))$; Javanmard and Nazerzadeh (2016) propose a policy for a similar problem that achieves regret $O(s \log d \log T)$, where s represents the sparsity of the d features; and in Ban and Keskin (2021), the near-optimal policy achieves regret $O(s\sqrt{T})$. In their parametric frameworks, the dependence of the regret on d is rather mild—it does not appear on the exponent of T; Javanmard and Nazerzadeh (2016) and Keskin and Zeevi (2014) also provide methods to deal with the sparse strucutre. In contrast, in our nonparametric formulation, the optimal rate of regret $O(T^{(2+d)/(4+d)})$ increases dramatically in d, making the problem significantly harder to learn in high dimensions. This is similar to the nonparametric formulation in the network revenue-management problem (Besbes and Zeevi 2012), in which the dimension of the decision space is d and the optimal rate of regret is $O(T^{(2+d)/(3+d)})$.[2] From the literature, it seems that the dimension significantly complicates the learning problem in a nonparametric formulation.

This paper is also related to the vast literature studying multi-armed bandit problems. See Cesa-Bianchi and Lugosi (2006) and Bubeck and Cesa-Bianchi (2012) for a comprehensive survey. The classic multi-armed bandit problem involves finite arms, and the algorithms (Agrawal and Goyal 2012, Kuleshov and Precup 2014) cannot be applied directly to our setting. Recently, there has been a stream of literature studying the so-called continuum-armed-bandit problems (Agrawal 1995, Kleinberg 2005, Auer et al. 2007, Kleinberg et al. 2008, Bubeck et al. 2011), in which there are infinite number of arms (decisions). Although there is no contextual information in those papers, Kleinberg et al. (2008) and Bubeck et al. (2011) have developed algorithms based on a similar idea to decision trees, because the potential arms form a high-dimensional space.

For multi-armed bandit problems with contextual information, parametric and regression-based algorithms have been proposed in, for example, Goldenshluger and Zeevi (2013) and Bastani and Bayati (2020). Our paper is related to the literature studying contextual multi-armed bandit problems in a nonparametric framework: Yang and Zhu (2002), Langford and Zhang (2008), Rigollet and Zeevi (2010), Perchet and Rigollet (2013), Slivkins (2014), and Elmachtoub et al. (2017). The analysis builds upon the idea of adaptive binning in Perchet and Rigollet (2013). Our algorithm is designed for continuous decisions. In fact, applying the algorithm in Perchet and Rigollet (2013) designed for discrete decisions to our problem with simple discretization may result in worse-than-optimal regret. In terms of the formulation and the rate of regret, this paper is closely related to Slivkins (2014). Slivkins (2014) investigates a more general model, in which the decision p can be a vector, and assumes that $f(\boldsymbol{x}, p)$ is Lipschitz-continuous. The optimal rate of regret in this setting has been shown to be $T^{(\mathbf{1}+d_x+d_p)/(2+d_x+d_p)}$ in previous works, where d_x and d_p are the dimensions of $\boldsymbol{x}$ and p, respectively. Slivkins (2014) introduces an adaptive zooming algorithm and uses the covering dimension of the space $(\boldsymbol{x}, p)$ in the analysis. The extension accommodates more general spaces of $(\boldsymbol{x}, p)$ than the Euclidean space. For Euclidean spaces, the algorithm recovers the optimal rate of regret $T^{(\mathbf{1}+d_x+d_p)/(2+d_x+d_p)}$. In our setting, letting $d_x = d$ and $d_p = 1$ leads to the rate $T^{(2+d)/(3+d)}$.[3] Motivated by personalized pricing, we impose additional assumptions on $f(\boldsymbol{x}, p)$ (smoothness and local concavity; see Assumption 3) and improve the optimal rate to $T^{(\mathbf{2}+d)/(\mathbf{4}+d)}$ as a result. The additional assumption may act as a special margin condition (Tsybakov 2004, Rigollet and Zeevi 2010, Perchet and Rigollet 2013) and affects the optimal rate. It is unclear whether the algorithm in Slivkins (2014) could be adapted to accommodate the additional assumptions and achieve the improved rate. The design of our algorithm is based on adaptively partitioning the covariate space into rectangular bins, rather than overlapping balls, as in Slivkins (2014).

2. Problem Formulation

Suppose the personalized demand function (purchasing probability) is $d(\boldsymbol{x}, p) \in [\mathbf{0}, \mathbf{1}]$, where $\boldsymbol{x} \in [\mathbf{0}, \mathbf{1})^d$ is the observed feature vector, or covariate, summarizing the customer's personalized information, and p is the price set by the firm. Because the purchasing event is a Bernoulli random variable with success rate $d(\boldsymbol{x}, p)$, when the firm sets price p for a customer of covariate $\boldsymbol{x}$ (henceforth abbreviated to customer $\boldsymbol{x}$), the expected revenue is thus $f(\boldsymbol{x}, p) \triangleq pd(\boldsymbol{x}, p)$. If the firm knew $d(\boldsymbol{x}, p)$, or equivalently, $f(\boldsymbol{x}, p)$, then it would set $p^*(\boldsymbol{x}) \triangleq \arg\max_{p \in [\mathbf{0},\mathbf{1}]} f(\boldsymbol{x}, p)$. Denote the optimal expected revenue from customer $\boldsymbol{x}$ by $f^*(\boldsymbol{x}) \triangleq \max_{p \in [\mathbf{0},\mathbf{1}]} \times f(\boldsymbol{x}, p)$. We will be primarily dealing with the expected revenue $f(\boldsymbol{x}, p)$ instead of the personalized demand $d(\boldsymbol{x}, p)$.

Initially, neither $d(\boldsymbol{x}, p)$ nor $f(\boldsymbol{x}, p)$ is known to the firm. In period $t \in \{\mathbf{1}, \mathbf{2}, \ldots, T\}$, a customer arrives with covariate $\boldsymbol{X}_t$. Upon observing $\boldsymbol{X}_t$, the firm sets a price p_t. The revenue earned in period t is denoted by Z_t, where Z_t/p_t is a Bernoulli random variable with mean $d(\boldsymbol{X}_t, p_t)$, independent of everything else. The objective of the firm is to design a pricing policy to maximize the total revenue over the horizon $\sum_{t=1}^T \mathsf{E}[Z_t] = \sum_{t=1}^T \mathsf{E}[f(\boldsymbol{X}_t, p_t)]$. Note that p_t itself is likely to be random, even though the firm is not adopting a randomized policy. This is because the pricing decision made in period t may depend on the observed customers, set prices, and earned revenues in the previous periods. That is, $p_t = \pi_t(\boldsymbol{X}_1, p_1, Z_1, \ldots, \boldsymbol{X}_{t-1}, p_{t-1}, Z_{t-1}, \boldsymbol{X}_t)$. Formally, we refer to π as the *pricing policy* that determines how p_t depends on the past information. We also denote $\mathcal{F}_t \triangleq \sigma(\boldsymbol{X}_1, p_1, Z_1, \ldots, \boldsymbol{X}_t, p_t, Z_t)$.

2.1. Regret

To measure the performance of a pricing policy, it is standard in the literature to benchmark it against the so-called *clairvoyant* policy and study the *regret*. Suppose $f(\boldsymbol{x}, p)$ is known to a clairvoyant firm. The optimal pricing policy is rather straightforward for a clairvoyant: Having observed customer $\boldsymbol{X}_t$, set $p^*(\boldsymbol{X}_t)$ in period t and earn a random revenue with mean $f^*(\boldsymbol{X}_t)$.

For the firm, the expected revenue in period t cannot exceed that of the clairvoyant: $f(\boldsymbol{X}_t, p_t) \le f^*(\boldsymbol{X}_t)$. Thus, we define the regret of a pricing policy π to be the revenue gap

$$R_\pi(T) = \sum_{t=1}^{T} \mathsf{E}\big[\big(f^*(\boldsymbol{X}_t) - f\big(\boldsymbol{X}_t, p_t\big)\big)\big].$$

In period t, the expectation is taken with respect to the distribution of $\boldsymbol{X}_t$ as well as p_t, which itself depends on $\mathcal{F}_{t-1}$ and $\boldsymbol{X}_t$. Our goal is to design a policy π that achieves small $R_\pi(T)$ when $T \to \infty$.

However, because $R_\pi(T)$ also depends on the unknown function f, we require the designed policy to perform well for a family $\mathcal{C}$ of function—that is, we seek for optimal policies in terms of the minimax regret

$$\inf_{\pi} \sup_{f \in \mathcal{C}} R_\pi(T).$$

Although it is usually impossible to find the exact policy that achieves the minimax regret, we focus on proposing a policy whose regret is at least comparable to (of the same order as) the minimax regret asymptotically when $T \to \infty$.

In this paper, we study functions f that cannot be parametrized. It implies that the family $\mathcal{C}$ is much larger than parametric families: It includes all the functions that satisfy some mild assumptions presented in the next section. In other words, the worst-case scenario can potentially be much worse than a parametric family. As a result, the achievable minimax regret is also higher.

2.2. Assumptions

In this section, we formally provide a set of assumptions that $f \in \mathcal{C}$ and the stochastic process has to satisfy and their justifications.

Assumption 1. *The covariates $\boldsymbol{X}_t$ are independent and identically distributed (i.i.d.) for $t = 1, \ldots, T$. Given $\boldsymbol{X}_t$ and p_t, the revenue Z_t is independent of everything else.*

Both i.i.d. covariates and independent noise structure are standard in the literature.

Assumption 2. *The functions $f(\cdot, p)$ and $f(\boldsymbol{x}, \cdot)$ are Lipschitz continuous given p and $\boldsymbol{x}$—that is, there exists $M_1 > 0$ such that $|f(\boldsymbol{x}_1, p) - f(\boldsymbol{x}_2, p)| \le M_1\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$ and $|f(\boldsymbol{x}, p_1) - f(\boldsymbol{x}, p_2)| \le M_1|p_1 - p_2|$ for all $\boldsymbol{x}_i$ and p_i $(i = 1, 2)$ in the domain.*

This assumption is equivalent to $|f(\boldsymbol{x}_1, p_1) - f(\boldsymbol{x}_2, p_2)| \le M_1(\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 + |p_1 - p_2|)$. Lipschitz continuity is a common assumption in the learning literature. In personalized pricing, it implies that the expected revenues are close if the firm charges similar prices for two customers with similar covariates. If this assumption fails, then the historical sales data of a certain type of customer are not informative for a new customer with almost identical background, and learning is virtually impossible.

To introduce the next assumption, consider any hyper-rectangle $B \subset [0, 1)^d$, including a singleton $B = \{\boldsymbol{x}\}$. Define $f_B(p) \triangleq \mathsf{E}[f(\boldsymbol{X}, p) | \boldsymbol{X} \in B]$ for $p \in [\boldsymbol{0}, \boldsymbol{1}]$. Clearly, $f_B(p)$ is the expected revenue when charging p for a customer that is sampled from a subset B.

Assumption 3. *We assume that for any B:*

1. *The function $f_B(p)$ has a unique maximizer $p^*(B) \in [0, 1]$. Moreover, there exist uniform constants $M_2, M_3 > 0$ such that for all $p \in [0, 1]$, $M_2(p^*(B) - p)^2 \le f_B(p^*(B)) - f_B(p) \le M_3(p^*(B) - p)^2$.*
2. *The maximizer $p^*(B)$ is inside the interval $[\inf\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in B\}, \sup\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in B\}]$.*
3. *Let d_B be the diameter of B. Then, there exists a uniform constant $M_4 > 0$ such that $\sup\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in B\} - \inf\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in B\} \le M_4 d_B$.*

This assumption is quite different from those in the setting without covariates (Besbes and Zeevi 2009, Wang et al. 2014, Lei et al. 2017) or the parametric setting (Qiang and Bayati 2016, Ban and Keskin 2021). To explain the intuition of $f_B(p)$, suppose the firm only observes $\mathbb{I}_{\{\boldsymbol{X} \in \boldsymbol{B}\}}$, but not the exact value of $\boldsymbol{X}$, and, thus, cannot apply personalized pricing for customers in B. In this case, the learning objective is the expected revenue $f_B(p)$, and the clairvoyant policy that has the knowledge of $f(\boldsymbol{x}, p)$ is to set $p = p^*(B)$. This class of learning problems are important subroutines of the algorithm we propose, and Assumption 3 guarantees that they can be effectively learned.

For part (1) of Assumption 3, we have

Proposition 1. *If $f_B(p)$ is continuous for $p \in [0, 1]$, and twice-differentiable in an open interval containing the unique global maximizer $p^*(B)$ with $f''_B(p^*(B)) < 0$, then part (1) of Assumption 3 holds.*

Therefore, part (1) states that $f_B(p)$ is smooth and locally concave around the maximum. If B is a singleton, then it can be viewed as a weaker version of the concavity assumption in Wang et al. (2014)

and Lei et al. (2017)—that is, $0 > a > f''(p) > b$ for all p in their no-covariate setting. As a result, if $f_B(p)$ is the revenue function of linear or exponential demand, then part (1) is satisfied automatically.

Part (2) of Assumption 3 prevents the following scenario: If the optimal price for the aggregate demand of customers $\boldsymbol{x} \in B$, $p^*(B)$, is far from the optimal personalized pricing $p^*(\boldsymbol{x})$, then collecting more information for $f_B(p)$ does not help to improve the pricing decision for any individual customer $\boldsymbol{x} \in B$. Such an obstacle may lead to failure to learn and is thus ruled out by the assumption. Similar types of assumptions have been imposed in other applications of revenue management. For example, proposition 1.16 in Gallego et al. (2018) provides conditions under which the optimal price of the aggregated market lies in the convex hull formed by the optimal prices of each market segment when discriminatory pricing is allowed.

Part (3) imposes a continuity condition for the optimal price. It is equivalent to, for example, some form of continuous differentiability of $f(\boldsymbol{x}, p)$, because $p^*(\boldsymbol{x})$ solves the implicit function from the first-order condition $f_p(\boldsymbol{x}, p) = \mathbf{0}$.

Remark 1. Assumption 2 is a variant of similar assumptions adopted in the literature. Assumption 3, although appearing nonstandard, is also satisfied by the parametric families studied by previous works. We give a few examples that satisfy Assumption 3.

- Dynamic pricing with linear covariate (Qiang and Bayati, 2016): If $f(\boldsymbol{x}, p) = p(\boldsymbol{\theta}^T\boldsymbol{x} - \boldsymbol{\alpha}p)$, then $f_B(p) = p(\boldsymbol{\theta}^T \mathsf{E}[\boldsymbol{X}|\boldsymbol{X} \in B] - \boldsymbol{\alpha}p)$ and $p^*(B) = \boldsymbol{\theta}^T \mathsf{E}[\boldsymbol{X}|\boldsymbol{X} \in B]/\mathbf{2}\boldsymbol{\alpha}$.
- Separable function: Consider $f(\boldsymbol{x}, p) = \sum_{i=\mathbf{1}}^k g_i(\boldsymbol{x})h_i(p)$. Then, $f_B(p) = \sum_{i=1}^k \mathsf{E}[g_i(\boldsymbol{X})|\boldsymbol{X} \in B]h_i(p)$. If $h_i(p)$ are concave functions and $g_i(\boldsymbol{x})$ are positive, then we may be able to solve the unique maximizer $p^*(B) = \mathsf{E}[g(\boldsymbol{X})|\boldsymbol{X} \in B]$ for some continuous function g.
- Localized functions: The covariate only plays a role in a subset $B_0 \subset [0, 1)^d$. See Section 5 for a concrete example.

As we shall see in Sections 4 and 5, the optimal rate of regret under Assumption 3 is $T^{(2+d)/(4+d)}$, in contrast to $T^{(2+d)/(3+d)}$ without it (taking $d_Y = 1$ in equation (3) of Slivkins 2014). Technically, we suspect that Assumption 3 plays a similar rule to the margin condition in the contextual bandit literature (Tsybakov 2004, Goldenshluger and Zeevi 2009, Rigollet and Zeevi 2010, Perchet and Rigollet 2013). However, because of the continuous decision variable studied in this paper, the margin condition cannot be translated in a straightforward way. It remains a future direction to present the assumption in a general form (the degree of smoothness such as the Hölder and Sobolev classes) and study how it affects the optimal rate of regret.

We summarize the information available to the firm. In the beginning of the horizon, the length of the horizon T, the dimension d, and the constants $\{M_i\}_{i=1}^4$ are revealed.[4] In period t, the price can also depend on $\mathcal{F}_{t-1}$ and $\boldsymbol{X}_t$.

3. The ABE Algorithm

We next present a set of preliminary concepts related to the *bins* of the covariate space and then introduce the proposed pricing policy: the Adaptive Binning and Exploration (ABE) algorithm.

3.1. Preliminary Concepts

Definition 1. A bin is a hyper-rectangle in the covariate space. More precisely, a bin is of the form

$$B = \{\boldsymbol{x} : a_i \le x_i < b_i,\ i = 1, \ldots, d\},$$

for $0 \le a_i < b_i \le 1, i = 1, \ldots, d$.

We can *split* a bin B by bisecting it in all the d dimensions to obtain 2^d *child* bins of B, all of equal size. For a bin B with boundaries a_i and b_i for $i = 1, \ldots, d$, its children are indexed by $\boldsymbol{i} \in \{\mathbf{0}, \mathbf{1}\}^d$ and have the form

$$B_{\boldsymbol{i}} = \left\{\boldsymbol{x} : a_j \le x_j < \frac{a_j + b_j}{2} \text{ if } \boldsymbol{i}_j = 0, \frac{a_j + b_j}{2} \le x_j < b_j \right.$$
$$\left. \times \text{ if } \boldsymbol{i}_j = 1,\ j = 1, \ldots, d\right\}.$$

Denote the set of child bins of B by $\boldsymbol{C}(\boldsymbol{B})$. Conversely, for any $B' \in \boldsymbol{C}(\boldsymbol{B})$, we refer to B as the *parent* bin of B', denoted by $P(B') = B$.

Our algorithm starts with a root bin $B_\emptyset \triangleq [0, 1)^d$, which contains all possible customers, and successively splits the bin as more data are collected. Therefore, any bin B produced during the process is the *offspring* of $B_\emptyset$—that is, $P^{(k)}(B) = B_\emptyset$ for some $k > 0$, where $P^{(k)}$ is the kth composition of the parent function. Equivalently, $B_\emptyset$ is an *ancestor* of B. For such a bin, we define its *level* to be k, denoted by $l(B) = k$. Conventionally, let $l(B_\emptyset) = 0$.

In the algorithm, we keep a dynamic partition $\mathcal{P}_t$ of the covariate space consisting of the offspring of $B_\emptyset$ in each period t. The partition is mutually exclusive and collectively exhaustive, so $B_i \cap B_j = \emptyset$ for $B_i, B_j \in \mathcal{P}_t$, and $\cup_{B_i \in \mathcal{P}_t} B_i = B_\emptyset$. Initially, $\mathcal{P}_0 = \{B_\emptyset\}$. In the algorithm, we gradually refine the partition; that is, each bin in $\mathcal{P}_{t+1}$ has an ancestor (or itself) in $\mathcal{P}_t$.

3.2. Intuition

The intuition behind the ABE algorithm is to use a partition $\mathcal{P}_t$ of the covariate space to aggregate customers in each period. It tries to find the optimal price for customers in each bin $B \in \mathcal{P}_t$—that is, $p^*(B)$ defined in Section 2.2. As $\mathcal{P}_t$ is refined dynamically—that is, $B \in \mathcal{P}_t$ becomes smaller—such aggregation is almost identical to personalized pricing.

To do that, we keep a set of discrete prices (referred to as the *decision set* hereafter) for each bin in the partition. The decision set consists of equally spaced grid points of a price interval associated with the bin. When a customer arrives with covariate $\boldsymbol{X}_t$ inside a bin B, a price is chosen successively in the decision set and charged for the customer. The realized revenue for this price is recorded. When a large number of customers are observed in B, the average revenue for each price p in the decision set is close to $f_B(p)$, which is defined as $\mathsf{E}[f(\boldsymbol{X},p)|\boldsymbol{X}\in B]$ in Section 2.2. Therefore, the *empirically optimal* price in the decision set is close to $p^*(B)$, with high confidence.

There are two potential pitfalls of this approach. First, the number of prices has an impact on the performance of the policy. If there are too many prices in a decision set, then, for a given number of customers observed in the bin, each price is experimented with for a relatively few times. As a result, the confidence interval for the associated average revenue is wide. On the other hand, if there are too few prices, then, inevitably, the decision set has low resolutions. That is, the optimal price in the set could still be far from the true maximizer $p^*(B)$ because of the discretization error. We have to select a proper size of the decision set to balance this trade-off.

Second, even if the optimal price $p^*(B)$ for the aggregate revenue in the bin is correctly identified, it may not be a strong indicator for $p^*(\boldsymbol{x})$ for a particular customer $\boldsymbol{x}\in B$. Indeed, $f_B(p)$ averages out all customers $\boldsymbol{X}\in B$, and the optimal price for an individual customer $\boldsymbol{x}$ could be very different. This obstacle, however, can be overcome as the size of B decreases, as implied by Assumption 3. In particular, parts (2) and (3) of the assumption guarantee that when B is small, $p^*(\boldsymbol{x})$ is concentrated within a neighborhood of $p^*(B)$, as long as $\boldsymbol{x}\in B$. The cost of using a smaller bin, however, is the lower frequency of observing a customer inside it.

To remedy the second pitfall, the algorithm adaptively refines the partition and decreases the size of the bins in $\mathcal{P}_t$ as t increases. When a bin $B\in\mathcal{P}_t$ is large, the aggregate optimal price $p^*(B)$ is not a strong indicator for $p^*(\boldsymbol{x})$, $\boldsymbol{x}\in B$. As a result, we only need a rough estimate and split the bin when a relatively small number of customers are observed in B. When a bin $B\in\mathcal{P}_t$ is small, the optimal price $p^*(B)$ provides a strong indicator for $p^*(\boldsymbol{x})$, $\boldsymbol{x}\in B$. Therefore, we gather large sales data from customers $\boldsymbol{X}\in B$ to explore the decision set and estimate $p^*(B)$ accurately, before it splits.

A crucial step in the algorithm is to determine what information to inherit when a bin is split into child bins. The ABE algorithm records the empirically optimal price in the decision set of the parent bin. In the child bins, we use this information and set up their decision sets centered at it. As explained above, when the parent bin (and, thus, the child bins) is large, its optimal price does not predict those of the child bins well. Therefore, the algorithm sets up conservative decision sets for the child bins—that is, they have wide intervals. On the other hand, when the parent bin is small, its optimal price provides an accurate indicator for those of the child bins. Thus, the algorithm constructs decision sets with narrow ranges for the child bins around the empirically-optimal price inherited.

Algorithm 1. Adaptive Binning and Exploration (ABE)

Input: T, d
Constants: M_1, M_2, M_3, M_4
Parameters: K; Δ_k, n_k, N_k for $k=0,\dots,K$
Initialize: partition $\mathcal{P}\leftarrow\{B_\emptyset\}$, $p_l^{B_\emptyset}\leftarrow 0$, $p_u^{B_\emptyset}\leftarrow 1$, $\delta_{B_\emptyset}\leftarrow 1/(N_0-1)$, $\bar{Y}_{B,j}, N_{B_\emptyset,j}\leftarrow 0$ for $j=0,\dots,N_0-1$
for $t=1$ to T **do**
  Observe $\boldsymbol{X}_t$
  $B\leftarrow\{B\in\mathcal{P}:\boldsymbol{X}_t\in B\}$ ▷ The bin in the partition that $\boldsymbol{X}_t$ belongs to
  $k\leftarrow l(B)$, $N(B)\leftarrow N(B)+1$ ▷ Determine the level and update the number of customers observed in B
  if $k<K$ **then** ▷ If not reaching the maximal level K
    if $N(B)<n_k$ **then** ▷ If not enough data observed in B
      $j\leftarrow N(B)-1 \pmod{N_k}$ ▷ Apply the jth price in the decision set
      $p_t\leftarrow p_l^B+j\delta_B$; apply p_t and observe Z_t
      $\bar{Y}_{B,j}\leftarrow\frac{1}{N_{B,j}+1}(N_{B,j}\bar{Y}_{B,j}+Z_t)$, $N_{B,j}\leftarrow N_{B,j}+1$
    else ▷ If sufficient data observed in B
      $j^*\in\arg\max_{j\in\{0,1,\dots,N_k-1\}}\{\bar{Y}_{B,j}\}$, $p^*\leftarrow p_l^B+j^*\delta_B$ ▷ Find the empirically optimal price; if there are multiple, choose any one of them
      $\mathcal{P}\leftarrow(\mathcal{P}\setminus B)\cup\boldsymbol{C}(\boldsymbol{B})$ ▷ Update the partition by removing B and adding its children
      for $B'\in\boldsymbol{C}(\boldsymbol{B})$ **do** ▷ Initialization for each child bin
        $N(B')\leftarrow 0$
        $p_l^{B'}\leftarrow\max\{0,p^*-\Delta_{k+1}/2\}$; $p_u^{B'}\leftarrow\min\{1,p^*+\Delta_{k+1}/2\}$ ▷ The range of the decision set
        $\delta_{B'}\leftarrow(p_u^{B'}-p_l^{B'})/(N_{k+1}-1)$ ▷ The grid size of the decision set
        $N_{B',j}, \bar{Y}_{B',j}\leftarrow 0$, for $j=0,\dots,N_{k+1}-1$ ▷ Initialize the average revenue and number of customers for each price
      end for
    end if
  else ▷ If reaching the maximal level
    $p_t\leftarrow(p_l^B+p_u^B)/2$
  end if
end for

3.3. Description of the Algorithm

In this section, we elaborate on the detailed steps of the ABE algorithm, shown in Algorithm 1.

The parameters for the algorithm include

1. K, the maximal level of the bins. When a bin is at level K, the algorithm no longer splits it and simply applies the median price of its decision set whenever a customer is observed in it.
2. Δ_k, the length of the interval that contains the decision set of level-k bins.
3. n_k, the maximal number of customers observed in a level-k bin in the partition. When n_k customers are observed, the bin splits.
4. N_k, the number of prices to explore in the decision set of level-k bins. The decision set of bin B consists of equally spaced grid points of an interval $[p_l^B, p_u^B]$, to be adaptively specified by the algorithm.

We initialize the partition to include only the root bin $B_\emptyset$ in step 4. Its decision set spans the whole interval $[0,1]$ with N_0 equally spaced grid points. That is, the jth price is $j\delta_{B_\emptyset} \triangleq j/(N_0-1)$ for $j=0,\ldots,N_0-1$. The initial average revenue and the number of customers that are charged the jth price are set to $\bar{Y}_{B_\emptyset,j} = N_{B_\emptyset,j} = 0$.

Suppose the partition is $\mathcal{P}_t$ at t and a customer $\boldsymbol{X}_t$ is observed (step 6). The algorithm determines the bin $B \in \mathcal{P}_t$, which the customer belongs to. The counter $N(B)$ records the number of customers already observed in B up to t when B is in the partition (step 8). If the level of B is $l(B)=k<K$ (i.e., B is not at the maximal level) and the number of customers observed in B is not sufficient (step 9 and step 10), then the algorithm has assigned a decision set to the bin in previous steps, namely, $\{p_l^B + j\delta_B\}$ for $j=0,\ldots,N_k-1$. There are N_k prices in the set, and they are equally spaced in the interval $[p_l^B, p_u^B]$. They are explored successively as new customers are observed in B (explore p_l^B for the first customer observed in B, $p_l^B+\delta_B$ for the second customer, $\ldots$, $p_l^B+(N_k-1)\delta_B$ for the N_kth customer, p_l^B again for the (N_k+1)th customer, etc.). Therefore, the algorithm charges price $p_t = p_l^B + j\delta_B$, where $j = N(B) - 1 (\mathrm{mod}\ N_k)$ for the $N(B)$th customer observed in B (step 11). Then, step 13 updates the average revenue and the number of customers for the jth price.

If the level of B is $l(B)=k<K$ and we have observed a sufficient number of customers in B (step 9 and step 14), then the algorithm splits B and replaces it by its 2^d child bins in the partition (step 16). For each child bin, step 18 to step 21 initialize the counter, the interval that encloses the decision set, the grid size of the decision set, and the average revenue/number of customers for each price in the decision set, respectively. In particular, to construct the decision set of a child bin, the algorithm first computes the empirically optimal price in the decision set of the parent bin B; that is, $j^* \in \arg\max_{j\in\{0,1,\ldots,N_k-1\}}\{\bar{Y}_{B,j}\}$ in step 15. Then, the algorithm creates an interval centered at this empirically optimal price with width Δ_{k+1}, properly cut off by the boundaries $[0,1]$. The decision set is then an equally spaced grid of the above interval (step 19 and step 20).

If the level of B is already K, then the algorithm simply charges the median price (step 25) repeatedly without further exploration. For such a bin, its size is sufficiently small, and the algorithm has narrowed the range of the decision set K times. The charged price is close enough to all $p^*(\boldsymbol{x})$, $\boldsymbol{x}\in B$, with high probability.

3.4. Choice of Parameters

We set $K = \lfloor \frac{\log(T)}{(d+4)\log(2)} \rfloor$, $\Delta_k = 2^{-k}\log(T)$, $N_k = \lceil \log(T) \rceil$, and

$$n_k = \max\left\{0, \left\lceil \frac{2^{4k+15}}{M_2^2 \log^3(T)} \left(\log(T) + \log(\log(T)) - (d+2)k\log(2)\right) \right\rceil \right\}.$$

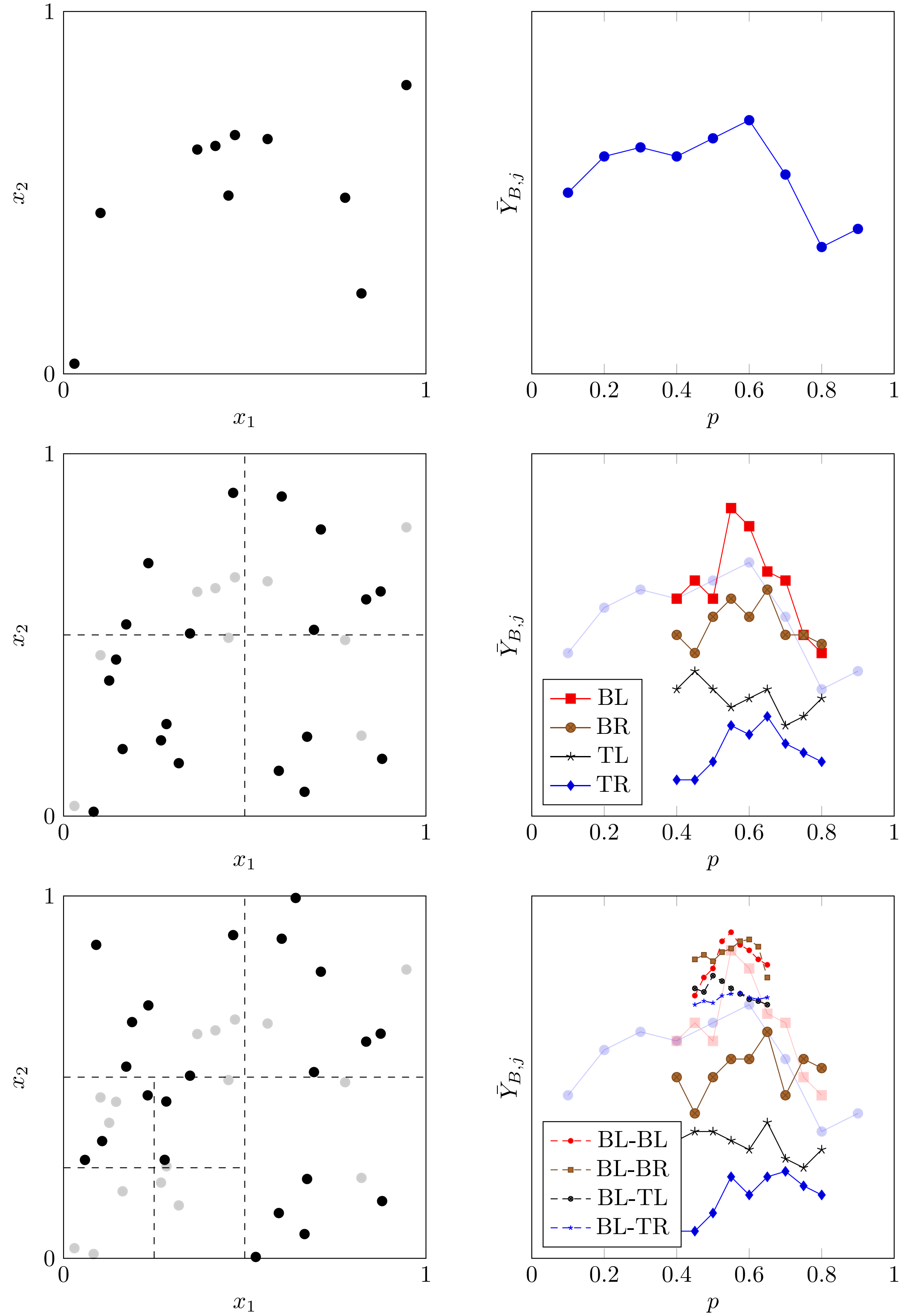
To give a sense of their magnitudes, the edge length of the bins at the maximal level is approximately $T^{-1/(d+4)}$. The range of the decision set (Δ_k) is proportional to the edge length of the bin (2^{-k}). The number of prices in a decision set is approximately $\log(T)$. Therefore, the grid size is $\delta_B \approx 2^{-k}$ for a level-k. The number of customers to observe in a level-k bin B is roughly $n_k \approx 2^{4k}/\log(T)^2$ before it splits. When k is small, n_k can be zero according to the expression. In this case, the algorithm immediately splits the bin without collecting any sales data in it.

3.5. A Schematic Illustration

We illustrate the key steps of the algorithm by an example with $d=2$. Figure 1 illustrates a possible outcome of the algorithm in periods $t_1 < t_2 < t_3$ (top panel, middle panel, and bottom panel, respectively). Up until period t_1, there is a single bin, and the covariates of observed customers $\boldsymbol{X}_t$ for $t \le t_1$ are illustrated in the top left panel. In this case, the decision set associated with the root bin is $p \in \{0.1, 0.2, \ldots, 0.9\}$, illustrated by the top right panel. The average revenue $\bar{Y}_{B,j}$ of each price is recorded, and $p^* = 0.6$ is the empirically optimal price. At t_1+1, a sufficient number of customers are observed, and step 14 is triggered in the algorithm. Therefore, the bin is split into four child bins.

From period t_1+1 to t_2, new customers are observed in each child bin (middle left panel). Note that the customers observed before t_1 in the parent bin are no longer used and are colored in gray. For each child bin (the bottom-left bin is abbreviated to BL, etc.), the average revenues for the prices in the decision sets are demonstrated in the middle right panel. The decision sets are centered at the empirically optimal price of

Figure 1. (Color online) A Schematic Illustration of the ABE Algorithm



their parent bin, which is $p^* = 0.6$ from the top right panel. They have narrower ranges and finer grids than that of the parent bin. At $t_2 + 1$, a sufficient number of customers are observed in BL, and it is split into four child bins.

From period $t_2 + 1$ to t_3, the partition consists of seven bins, as shown in the bottom left panel. The BR, TL, and TR bins keep observing customers and updating the average revenues because they have not collected sufficient data. Their status at t_3 is shown in the bottom panels. In the four newly created child bins of BL (the bottom-left bin of BL is abbreviated to BL-BL, etc.), the prices in the decision sets are used successively, and their average revenues are illustrated in the bottom right panel.

4. Regret Analysis: Upper Bound

To measure the performance of the ABE algorithm, we provide an upper bound for its regret.

Theorem 1. *For any function f satisfying Assumptions 2 and 3, the regret incurred by the ABE algorithm is bounded by*

$$R_{\pi_{ABE}}(T) \leq CT^{\frac{2+d}{4+d}} \log(T)^2,$$

for a constant $C > 0$ that is independent of T.

We provide a sketch of the proof here and present the details in the online appendix. In period t, if $\boldsymbol{X}_t \in B$ for a bin in the partition $B \in \mathcal{P}_t$, then the expected regret incurred by the ABE algorithm is $\mathsf{E}[(f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t, p_t))\mathbb{I}_{\{\boldsymbol{X}_t \in B, B \in \mathcal{P}_t\}}]$. Because the total regret simply sums up the above quantity over $t = 1, \ldots, T$ and all possible Bs, it suffices to focus on the regret for given t and B. Two possible scenarios can arise: (1) The optimal price of the aggregate demand in B—that is, $p^*(B)$—is inside the range of the decision set—that is, $p^*(B) \in [p_l^B, p_u^B]$ (step 19); and (2) the optimal price $p^*(B)$ is outside the range of the decision set.

Scenario 1 represents the regime where the algorithm is working "normally": Up until t, the algorithm has successfully narrowed the optimal price $p^*(B)$ (which provides a useful indicator for all $p^*(\boldsymbol{x})$, $\boldsymbol{x} \in B$ when B is small) down to $[p_l^B, p_u^B]$. By Assumption 3(1), the regret in this scenario can be decomposed into two terms:

$$f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t, p_t) \leq M_3(|p_t - p^*(B)| + |p^*(B) - p^*(\boldsymbol{X}_t)|)^2.$$

The first term can be bounded by the length of the interval $p_u^B - p_l^B$. The second term can be bounded by the size of B given $\boldsymbol{X}_t \in B$ by Assumption 3, (2) and (3). By the choice of parameters in Section 3.4, the length of the interval decreases as the bin size decreases. Therefore, both terms can be well controlled when the size of B is sufficiently small, or, equivalently, when the level $l(B)$ is sufficiently large. This is why a properly chosen n_k can guarantee that the algorithm spends little time for large bins and collects a large amount of data for small bins. When the bin level reaches K, the above two terms are small enough, and no more exploration is needed.

Scenario 2 represents the regime where the algorithm works "abnormally." In scenario 2, the difference $f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t, p_t)$ can no longer be controlled as in scenario 1 because p_t and $p^*(\boldsymbol{X}_t)$ can be far apart. To make things worse, $p^*(B) \notin [p_l^B, p_u^B]$ usually implies $p^*(B') \notin [p_l^{B'}, p_u^{B'}]$, where B' is a child of B. This is because (1) $p^*(B)$ is close to $p^*(B')$ for small B, and (2) $[p_l^{B'}, p_u^{B'}]$ is created around the empirically optimal price for B, and thus overlapping with $[p_l^B, p_u^B]$. Therefore, for any period s following t, the worst-case regret is $O(1)$ in that period if $\boldsymbol{X}_s \in B$ or its offspring.

To bound the regret in scenario 2, we have to bound the probability, which requires delicate analysis of the events. If scenario 2 occurs for B, then during the process that we sequentially split $B_\emptyset$ to obtain B, we can find an ancestor bin of B (which can be B itself) that scenario 2 happens for the first time along the "branch" from $B_\emptyset$ all the way down to B. More precisely, denoting the ancestor bin by B_a and its parent by $P(B_a)$, we have (1) $p^*(P(B_a))$ is inside $[p_l^{P(B_a)}, p_u^{P(B_a)}]$ (scenario 1); (2) after $P(B_a)$ is split, $p^*(B_a)$ is outside $[p_l^{B_a}, p_u^{B_a}]$ (scenario 2). Denote the empirically optimal price in the decision set of $P(B_a)$ by p^*. For such an event to occur, the center of the decision set of B_a, which is p^*, has to be at least $\Delta_{l(B_a)}/2$ away from $p^*(B_a)$.[5] Because of Assumption 3 and the choice of Δ_k, the distance between $p^*(P(B_a))$ and $p^*(B_a)$ is relatively small compared with $\Delta_{l(B_a)}$. Therefore, the empirically optimal price p^* must be far away from $p^*(P(B_a))$. The probability of such an event can be bounded by using classic concentration inequalities for sub-Gaussian random variables: The prices that are closer to $p^*(P(B_a))$ and, thus, have higher means turn out to generate lower average revenue than p^*; this event is extremely unlikely to happen when we have collected a large amount of sales data for each price in the decision set.

The total regret aggregates those in scenarios 1 and 2 for all possible combinations of B and B_a. It matches the quantity $O(\log(T)^2 T^{(2+d)/(4+d)})$ presented in the theorem.

5. Regret Analysis: Lower Bound

In this section, we show that the minimax regret is no lower than $cT^{(2+d)/(4+d)}$ for some constant c. Combining with the last section, we conclude that no nonanticipating policy does better than the ABE algorithm in terms of the order of magnitude of the regret in T (neglecting logarithmic terms).

We first construct a family of functions that satisfy Assumptions 2 and 3. The functions in the family are selected to be "difficult" to distinguish. By doing so, we will prove that any policy has to spend a substantial amount of time exploring prices that generate low revenues, but help to differentiate the functions. Otherwise, the incapability to correctly identify the underlying function is costly in the long run. Therefore, unable to contain both sources of regret at the same time, no policy can achieve lower regret than the quantity stated in Theorem 2.

Before introducing the family of functions, we define ∂B to be the boundary of a convex set in $[0,1)^d$. Let $D(B_1, B_2)$ be the Euclidean distance between two sets B_1 and B_2. That is, $D(B_1, B_2) \triangleq \inf\{\|\mathbf{x_1} - \mathbf{x_2}\|_2 : \mathbf{x_1} \in B_1, \mathbf{x_2} \in B_2\}$. We allow B_1 or B_2 to be a singleton. To define $\mathcal{C}$, we partition the covariate space $[0,1)^d$ into M^d equally sized bins. That is, each bin has the following form: For $(k_1, \ldots, k_d) \in \{1, \ldots, M\}^d$,

$$\left\{\mathbf{x} : \frac{k_i - 1}{M} \le x_i < \frac{k_i}{M}, \quad \forall\, i = 1, \ldots, d\right\}.$$

We number those bins by $1, \ldots, M^d$ in an arbitrary order—that is, $B_1, \ldots, B_{M^d}$. Each function $f(\mathbf{x}, p) \in \mathcal{C}$ is indexed by a tuple $w \in \{\mathbf{0}, \mathbf{1}\}^{M^d}$, whose jth index determines the behavior of $f_w(\mathbf{x}, p)$ in B_j. More precisely, for $\mathbf{x} \in B_j$, the personalized demand function is

$$d_w(\mathbf{x}, p) = \begin{cases} \frac{2}{3} - \frac{p}{2} & w_j = 0 \\ \frac{2}{3} - \frac{p}{2} + \left(\frac{1}{3} - \frac{p}{2}\right) D\left(\mathbf{x}, \partial B_j\right) & w_j = 1 \end{cases},$$

and, thus,

$$f_w(\mathbf{x}, p) = \begin{cases} p\left(\frac{2}{3} - \frac{p}{2}\right) & w_j = 0 \\ p\left(\frac{2}{3} - \frac{p}{2} + \left(\frac{1}{3} - \frac{p}{2}\right) D\left(\mathbf{x}, \partial B_j\right)\right) & w_j = 1 \end{cases}.$$

The optimal personalized price for customer $\mathbf{x} \in B_j$ is $p^*(\mathbf{x}) = \mathbf{2/3}$ if $w_j = 0$ and $p^*(\mathbf{x}) = \frac{\mathbf{2} + D(\mathbf{x}, \partial B_j)}{3(1 + D(\mathbf{x}, \partial B_j))}$ if $w_j = 1$.

The construction of $\mathcal{C}$ follows a similar idea to Rigollet and Zeevi (2010). For a given $f_w \in \mathcal{C}$, we can always find another function $f_{w'} \in \mathcal{C}$ that only differs from f in a single bin B_j by setting w' to be equal to w except for the jth index. The firm can only rely on the covariates generated in B_j to distinguish between f_w and $f_{w'}$. For small bins (i.e., large M), this is particularly costly because there are only a tiny fraction of customers observed in a particular bin, and the difference $|f_w - f_{w'}| = p(1/3 - p/2)D(\mathbf{x}, \partial B_j)$ becomes tenuous. It also requires p to be far from 2/3 to detect the difference, which happens to be the optimal price when $w_j = 0$. This makes the exploration/exploitation trade-off hard to balance. Now, a policy has to carry out the task for M^d bins—that is, distinguishing the underlying function f_w with M^d tuples that only differ from w in one index. The cost is inevitable and adds to the lower bound of the regret. Moreover, we assume the customers $\mathbf{X}$ are uniformly distributed in $[\mathbf{0}, \mathbf{1})^d$.

In the online appendix, we show that the constructed $\mathcal{C}$ satisfies all the assumptions. The main theorem below shows the lower bound for the regret.

Theorem 2. *For the constructed $\mathcal{C}$, any nonanticipating policy π has regret*

$$\sup_{f \in \mathcal{C}} R_\pi(T) \ge cT^{\frac{2+d}{4+d}},$$

for a constant $c > 0$.

6. Future Research

As shown in Theorems 1 and 2, the best achievable regret of the problem is of order $T^{(2+d)/(4+d)}$. As a result, the knowledge of the sparsity structure of the covariate is essential in designing pricing policies. More precisely, the provided customer covariate is of dimension d, while the personalized demand $d(\mathbf{X}, p)$ may only depend on d' entries of the covariate, where $d' \ll d$. In this case, being able to identify the d' entries out of d significantly decreases the incurred regret from $T^{(2+d)/(4+d)}$ to $T^{(2+d')/(4+d')}$. Indeed, in the ABE algorithm, if the sparsity structure is known, then a bin is split into $2^{d'}$ instead of 2^d child bins. It pools the observations that only differ in the dimensions corresponding to the redundant covariates so that more observations are available in a bin, and, thus, substantially reduces the exploration cost. An important research question is then whether it is possible to design a binning algorithm that selects one dimension and the position to split, based on a certain criterion, like regression/classification decision trees (Hastie et al. 2001). This may significantly improve the regret in the presence of sparse covariates.

Endnotes

[1] Perchet and Rigollet (2013) study discrete decision variables (multiarmed bandit).

[2] It is shown in Chen and Gallego (2018) that if the number of inventory constraints $\ll d$, then learning the dual variables may effectively reduce the problem dimension.

[3] Applying our assumptions and following equation (8) of Slivkins (2014) gives $d_x + d_p = d + 1/2$, and the regret improves to $T^{(1.5+d)/(2.5+d)}$.

[4] In fact, only M_2 is needed in the algorithm.

[5] Recall that $p_u^{B_a} - p_l^{B_a} = \Delta_{l(B_a)}$.

References

Agrawal R (1995) The continuum-armed bandit problem. *SIAM J. Control Optim.* 33(6):1926–1951.

Agrawal S, Goyal N (2012) Analysis of Thompson sampling for the multi-armed bandit problem. *JMLR Workshop Conf Proc.* 23: 39.1–39.26.

Araman VF, Caldentey R (2009) Dynamic pricing for nonperishable products with demand learning. *Oper. Res.* 57(5):1169–1188.

Auer P, Ortner R, Szepesvári C (2007) *Improved Rates for the Stochastic Continuum-Armed Bandit Problem* (Springer, Berlin), 454–468.

Ban G, Keskin NB (2021) Personalized dynamic pricing with machine learning: High dimensional features and heterogeneous elasticity. *Management Sci.* Forthcoming.

Bastani H, Bayati M (2020) Online decision-making with high-dimensional covariates. *Oper. Res.* 68(1):276–294.

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

Besbes O, Zeevi A (2012) Blind network revenue management. *Oper. Res.* 60(6):1537–1550.

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

Bubeck, S. and N. Cesa-Bianchi (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations Trends® Machine Learn.* 5(1):1–122.

Bubeck S, Munos R, Stoltz G, Szepesvári C (2011) X-armed bandits. *J. Machine Learn. Res.* 12(May):1655–1695.

Cesa-Bianchi N, Lugosi G (2006) *Prediction, Learning, and Games* (Cambridge University Press, Cambridge, UK).

Chen N, Gallego G (2018) A primal-dual learning algorithm for personalized dynamic pricing with an inventory constraint. Working paper, University of Toronto, Mississauga, Canada.

Cheung WC, Simchi-Levi D, Wang H (2017) Dynamic pricing and demand learning with limited price experimentation. *Oper. Res.* 65(6):1722–1731.

Cohen MC, Lobel I, Paes Leme R (2020) Feature-based dynamic pricing. *Management Sci.* 66(11):4921—5484.

den Boer AV (2015) Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys Oper. Res. Management Sci.* 20(1):1–18.

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

Elmachtoub AN, McNellis R, Oh S, Petrik M (2017) A practical method for solving contextual bandit problems using decision trees. Working paper, Columbia University, New York.

Farias VF, Van Roy B (2010) Dynamic pricing with a prior on market response. *Oper. Res.* 58(1):16–29.

Gallego G, Topaloglu H (2018) Revenue Management and Pricing Analytics, *International Series in Operations Research & Management Science*, vol. 279 (Springer, New York).

Goldenshluger A, Zeevi A (2009) Woodroofe's one-armed bandit problem revisited. *Ann. Appl. Probab.* 19(4):1603–1633.

Goldenshluger A, Zeevi A (2013) A linear response bandit problem. *Stochastic Systems* 3(1):230–261.

Hastie T, Tibshirani R, Friedman J (2001) *The Elements of Statistical Learning*, Springer Series in Statistics (Springer, New York).

Javanmard A, Nazerzadeh H (2016) Dynamic pricing in high-dimensions. *J. Machine Learn. Res.* 20(1):315–363.

Keskin NB, Zeevi A (2014) Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Oper. Res.* 62(5):1142–1167.

Kleinberg R, Slivkins A, Upfal E (2008) Multi-armed bandits in metric spaces. *Proc. 40th Annu. ACM Sympos. Theory Comput.* (Association for Computing Machinery, New York), 681–690.

Kleinberg RD (2005) Nearly tight bounds for the continuum-armed bandit problem. Saul LK, Weiss Y, Bottou L, eds. *Proc. 17th Internat. Conf. Neural Inform. Processing Systems* (MIT Press, Cambridge, MA), 697–704.

Kuleshov V, Precup D (2014) Algorithms for multi-armed bandit problems. Working paper, McGill University, Montreal.

Langford J, Zhang T (2008) The epoch-greedy algorithm for multi-armed bandits with side information. Platt JC, Koller D, Singer Y, Roweis ST, eds. *Advances in Neural Information Processing Systems*, vol. 20 (Neural Information Processing Systems Foundation, Red Hook, NY), 817–824.

Lei Y, Jasin S, Sinha A (2017) Near-optimal bisection search for nonparametric dynamic pricing with inventory constraint. Working paper, Queen's University, Kingston, Canada.

Nambiar M, Simchi-Levi D, Wang H (2016) Dynamic learning and pricing with model misspecification. *Management Sci.* 65(11): 4980–5000.

Perchet V, Rigollet P (2013) The multi-armed bandit problem with covariates. *Ann. Statist.* 41(2):693–721.

Qiang S, Bayati M (2016) Dynamic pricing with demand covariates. Working paper, Stanford University, Stanford, CA.

Rigollet P, Zeevi A (2010) Nonparametric bandits with covariates. Kalai AT, Mohri M, eds. *Proc. 23rd Conf. Learn. Theory (COLT)* (Association for Computing Machinery, New York), 54–66.

Slivkins A (2014) Contextual bandits with similarity information. *J. Machine Learn. Res.* 15(1):2533–2568.

Tsybakov AB (2004) Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* 32(1):135–166.

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

Yang Y, Zhu D (2002) Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Ann. Statist.* 30(1):100–121.

Ningyuan Chen is currently an assistant professor of the Department of Management at the University of Toronto, Mississauga. He is also cross-appointed at the Rotman School of Management, University of Toronto. Before joining the University of Toronto, he was an assistant professor of HKUST. He received his PhD from the Industrial Engineering and Operations Research department at Columbia University.

Guillermo Gallego is the department head of Industrial Engineering and Decision Analytics, the Crown Worldwide Professor of Engineering at HKUST, and the Liu Family Emeritus Professor at Columbia University. Prof Gallego's research interests are in Discrete Choice Models, Assortment Optimization, and Pricing Analytics with applications to Revenue Management, Electronic Commerce, and Inventory Theory.