Policy Optimization Using Semiparametric Models for Dynamic Pricing

Abstract

In this paper, we study the contextual dynamic pricing problem where the market value of a product is linear in its observed features plus some market noise. Products are sold one at a time, and only a binary response indicating success or failure of a sale is observed. Our model setting is similar to Javanmard and Nazerzadeh [2019] except that we expand the demand curve to a semiparametric model and learn dynamically both parametric and nonparametric components. We propose a dynamic statistical learning and decision making policy that minimizes regret (maximizes revenue) by combining semiparametric estimation for a generalized linear model with unknown link and online decision making. Under mild conditions, for a market noise c.d.f. $F(\cdot)$ with m-th order derivative ($m \geqslant 2$), our policy achieves a regret upper bound of $\widetilde{\mathcal{O}}_d(T^{\frac{2m+1}{4m-1}})$, where T is the time horizon and $\widetilde{\mathcal{O}}_d$ is the order hiding logarithmic terms and the feature dimension d. The upper bound is further reduced to $\widetilde{\mathcal{O}}_d(\sqrt{T})$ if F is super smooth. These upper bounds are close to $\Omega(\sqrt{T})$, the lower bound where F belongs to a parametric class. We further generalize these results to the case with dynamic dependent product features under the strong mixing condition.

Keyword: Contextual dynamic pricing, generalized linear model with unknown link, policy optimization, non-parametric statistics.

1 Introduction

Dynamic pricing is the study of determining and adjusting the selling prices of products over time based on statistical learning and policy optimization. As an integral part of revenue management, it has wide applications to various industries. Research on dynamic pricing has spanned across the fields of statistics, machine learning, economics, and operations research [den Boer, 2015, Wei and Zhang, 2018, Misic and Perakis, 2020]. In general, a good pricing strategy often involves good statistical learning of the demand function as well as revenue optimization over time.

Recent works particularly focus on feature-based (or contextual) pricing models, where the market value of a product as well as the pricing strategy depend on some observable features of the product [Javanmard and Nazerzadeh, 2019, Ban and Keskin, 2020]. Given the product features (covariates) available through the massive real-time data in online platforms today, feature-based pricing models take product heterogeneity into account, which enable customized pricing for products.

In this work, we consider the following dynamic pricing problem: We assume that a seller sells one product at each time $t = 1, \cdots, T$. Each product is attached with a known feature vector $\mathbf{x}_t \in \mathbb{R}^d$. In addition, the product's market value v_t is linear in the features plus some i.i.d. market noise z_t with an *unknown* cumulative distribution $F(\cdot)$:

$$v_t = \boldsymbol{\theta}_0^\top \widetilde{\mathbf{x}}_t + z_t, \qquad z_t \sim F.$$

Here $\widetilde{\mathbf{x}}_t = (\mathbf{x}_t^\top, 1)^\top$ and $\boldsymbol{\theta}_0$ is some unknown parameter. The customer makes an independent purchase decision for each product depending on whether the seller's posted price p_t is higher than the market value v_t, after which the revenue is collected. In this case, the demand curve $P(v_t \geqslant p_t)$ actually depends on both the parameter $\boldsymbol{\theta}_0$ as well as the distribution of z_t, which admits a semi-parametric form. They need to be learned or estimated dynamically from the observed binary data that indicates whether a sale is successful. Under this setting, we propose a policy which utilizes semi-parametric estimation techniques to achieve a low regret. In particular, under mild regularity conditions, if the c.d.f. $F(\cdot)$ of z_t has m^{th} derivative, the regret over a time horizon T is upper bounded by $\mathcal{O}((Td)^{\frac{2m+1}{4m-1}} \log T(1 + \log T/d))$, where d is the number of features. This result is further generalized to a setting where the product features $\mathbf{x}_t$ are not independent, as long as $\{\mathbf{x}_t\}_{t \geqslant 1}$ is a stationary series that satisfies certain β-mixing conditions. Moreover, when F is infinitely differentiable, the total regret can be upper bounded by $\widetilde{\mathcal{O}}((Td)^{\frac{1}{2}}(\log T)^{\frac{3}{2}+\frac{3}{2\alpha}}(\log(d+1)+\log T/d))$. This rate is the same as the parametric lower bound up to some logarithmic factors, i.e. where the distribution of z_t is generated from a parametric class.

1.1 Related Literatures

Our work contributes to the recent line of dynamic pricing literature as well as the growing literature on decision making with covariate information and contributes to kernel regression. Our work is also closely related to the non-parametric statistics literature. We'll briefly review the related works in the below.

Dynamic pricing. In the classical pricing models, one aims at maximizing the revenue over time by posting price sequentially while learning the underlying demand curve or market evaluation of a product. The demand curve is typically fixed over time, and falls into a known function class. Related literature includes Kleinberg and Leighton [2003], Rusmevichientong et al. [2006], Besbes and Zeevi [2009], Broder and Rusmevichientong [2012], Keskin and Zeevi [2014], den Boer and Zwart [2014], Wang et al. [2014], den Boer and Zwart [2015], Babaioff et al. [2015], Cesa-Bianchi et al. [2019], Chen et al. [2019]. As an example, Cesa-Bianchi et al. [2019] study the dynamic pricing problem where the buyer's valuation of a product is supported on a finite K unknown points, and the success of a sale is determined by comparing the valuation to the proposed price. Using a generalization of UCB algorithm, the authors achieve the regret with order $\mathcal{O}(K \log T)$. For a comprehensive survey on this topic, see den Boer [2015].

Recently, many papers have been focusing on contextual dynamic pricing, where product heterogeneity is taken into account when modeling the demand curve or market evaluation. A common and natural choice is to model the market value of the product at time t as a linear function of its features $\mathbf{x}_t$ plus some market noise z_t, i.e. $v_t = \boldsymbol{\theta}^\top \mathbf{x}_t + z_t$ where $\boldsymbol{\theta}$ is some unknown parameter [Qiang and Bayati, 2016, Javanmard, 2017, Miao et al., 2019, Javanmard and Nazerzadeh, 2019, Ban and Keskin, 2020, Wang et al., 2020a, Chen et al., 2021, Tang et al., 2020, Golrezaei et al., 2020]. Under this setting, for 'truthful' buyers whose decision is based on comparing v_t and offered price p_t, the demand curve can be expressed as a generalized linear model given feature covariates $\mathbf{x}_t$, where the link function is closely related to the distribution of the market noise z_t (see (3) for a detailed reasoning). Qiang and Bayati [2016] assume a linear model between the demand curve and the product features. They prove that the greedy iterative least squares (GILS) algorithm achieves a regret upper bound of $\mathcal{O}_d(\log T)$, where $\widetilde{\mathcal{O}}_d$ is the order that hides logarithmic terms and the dimensionality of feature d, and provide a matching lower bound under their setting. Miao et al. [2019] and Ban and Keskin [2020] consider a generalized linear model with known link, while Javanmard and Nazerzadeh [2019] and Wang et al. [2020a] study the same problem with high dimensional sparse parameters. The algorithms are usually a combination of statistical estimation procedures and online learning techniques. Depending on the setting, the optimal regret ranges from $\mathcal{O}_d(\log T)$ to $\widetilde{\mathcal{O}}_d(\sqrt{T})$. Other related works include Chen et al. [2021], Tang et al. [2020] where the authors explore certain differentially private policies under similar model setting; Golrezaei et al. [2020] where the authors consider the second price auction problem with multiple customers, each of which has his/her own product evaluation; and Javanmard [2017] where the parameter $\boldsymbol{\theta}$ in the generalized linear model changes through time.

In practice, however, the distribution of the market noise z_t is usually unknown to the seller. Thus, it might be desirable to only assume that the noise density falls into some general class. As will be discussed in §2, this leads to modeling the demand curve as a generalized linear model with unknown link, and will be our main focus in this paper. Compared to the previous setting, this setting is more challenging, and the related literature is sparse. Javanmard and Nazerzadeh [2019] propose a preliminary algorithm that achieves a regret upper bound of $\mathcal{O}_d(T)$. Golrezaei et al. [2019] consider a second price auction with reserve where there are more than one customers, each of whom has his/her individual parameters in their demand curve model, and the customer bids are available as additional information. The authors propose the NPAC-T/NPAC-S policy that achieves a regret $\widetilde{\mathcal{O}}_d(\sqrt{T})$. Golrezaei et al. [2020] also explore the second price auction and derive a regret upper bound of $\widetilde{\mathcal{O}}_d(T^{2/3})$ compared to a 'robust benchmark' where the price maximizes the revenue of the worst link function in the class. Shah et al. [2019] explore an alternative setting where the market value $v_t = \exp(\boldsymbol{\theta}^\top \mathbf{x}_t + z_t)$ and z_t has unknown distribution. By utilizing this specific structure, the authors propose the DEEP-C algorithm based on multi-arm bandit that has a regret upper bound of $\widetilde{\mathcal{O}}_d(\sqrt{T})$. The authors also propose some variants of the algorithm and study them via simulations. Recently, Luo et al. [2021] study a similar problem to ours, assuming a linear market valuation with unknown noise distribution. They provide a DIP policy that achieves regret $\mathcal{O}_d(T^{2/3} + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1 T)$, where $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1$ is the estimation accuracy of the parameter $\boldsymbol{\theta}$.

There are some literature studying other dynamic pricing algorithms [Amin et al., 2014, Cohen et al., 2016, Mao et al., 2018, Leme and Schneider, 2018, Nambiar et al., 2019, Anton and Alexey, 2020, Alexey, 2020, Ban and Keskin, 2020, Li and Zheng, 2020, Javanmard et al., 2020, Chen and Gallego, 2020, Liu et al., 2021]. For example, Mao et al. [2018] study a non-parametric dynamic pricing pricing where the market value is modeled as a general non-parametric function $f(\mathbf{x}_t)$, where $\mathbf{x}_t$ are the features. A binary feedback is similarly observed based on the comparison between $f(\mathbf{x}_t)$ and the proposed price. The authors apply a variation of midpoint algorithm and achieve a regret upper bound of $\mathcal{O}(T^{d/(d+1)})$ with d being the dimension of $\mathbf{x}_t$.

Semi-parametric and non-parametric statistical estimation. Our work is also closely related to estimation of the single index model, or the generalized linear model with an unknown link. Such model has been studied in the statistics and econometrics literature for decades, and has wide applications in fields like econometrics and finance [Powell et al., 1989, Ichimura, 1993, Hardle et al., 1993, Klein and Spady, 1993, Weisberg and Welsh, 1994, Mallick and Gelfand, 1994, Horowitz and Härdle, 1996, Carroll et al., 1997, Xia and Li, 1999, Delecroix et al., 2003, Fan and Li, 2004]. For a comprehensive summary of these works, please refer to McCulloch [2000], Györfi et al. [2002], Fan and Yao [2003], Ruppert et al. [2003], Tsybakov [2008], Horowitz [2012]. Various methods have been proposed to estimate the parametric part that achieves root-n consistency under certain conditions [Powell et al., 1989, Ichimura, 1993, Klein and Spady, 1993]. Carroll et al. [1997] study the generalized partial linear single index models, where the authors leverage local linear kernel regression with quasi-likelihood method to estimate both the parametric and non-parametric parts of the model. Xia and Li [1999] investigate in the single index coefficient model with strong-mixing features. Estimators with uniform convergence rate to the ground truth based on kernel regression are proposed.

Given a root-n consistent estimation of the coefficients, standard univariate non-parametric regression techniques can be used to estimate the non-parametric part of the single index model that achieves

ℓ_∞ consistency, which is necessary in deriving regret upper bounds. One common estimator is the Nadaraya-Watson estimator [Nadaraya, 1964, Watson, 1964]. Silverman [1978] and Mack and Silverman [1982] establish uniform convergence results for kernel density estimator and Nadaraya-Watson estimator for regression functions. In addition, Stone [1980, 1982] derive uniform convergence results for the more general local polynomial regression estimators. Masry [1996] prove similar results when the covariates satisfy strong-mixing conditions.

In this paper, we'll provide non-asymptotic error bounds for both coefficient estimation as well as the plug-in Nadaraya-Watson estimator in a uniform sense. These non-asymptotic bounds are useful for constructing regret bounds within a finite horizon.

1.2 Our Contributions

Our contributions are the following: First, compared to related works, our policy achieves a low regret with few assumptions on the market noise distribution and little additional information. Given $F \in \mathbb{C}^{(m)}$ where F is the c.d.f. of z_t, the regret over a time horizon T is upper bounded by $\widetilde{\mathcal{O}}((Td)^{\frac{2m+1}{4m-1}})$; If F is 'super smooth', the bound is further reduced to $\widetilde{\mathcal{O}}(\sqrt{Td})$, which is nearly the same regret order by assuming a parametric distribution for z_t as in Javanmard and Nazerzadeh [2019] where the s-sparsity on $\boldsymbol{\beta}_0$ is imposed. Table 1 illustrates the settings of our work as well as several related literatures. Golrezaei et al. [2020] choose a more 'conservative' regret by comparing to a benchmark policy which minimizes revenue with the worst demand function over the whole ambiguity function class. In contrast, our notation of regret is more standard and 'accurate' in that our benchmark policy knows the exact demand function given any product features. Shah et al. [2019] consider a log-linear relation between the market value and the covariates instead of a linear relation and derive a regret upper bound of $\widetilde{\mathcal{O}}(\sqrt{T}d^{11/4})$. Their algorithm based on multi-arm bandit has sub-optimal dependence on the dimension d in terms of both regret and complexity, and is quite difficult to implement under general conditions. Interestingly, the authors conjecture that under the linear settings, there is no policy that achieves an $\widetilde{\mathcal{O}}_d(\sqrt{T})$ regret. Our work partly answers their guess by providing a policy with a $\widetilde{\mathcal{O}}(\sqrt{Td})$ regret when the demand function is sufficiently smooth. Compared to the DIP policy in Luo et al. [2021] and its regret $\mathcal{O}_d(T^{2/3} + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1 T)$, we are more clear on how $\widehat{\boldsymbol{\theta}}$ are estimated

	Feature-based	Non-parametric noise	Regret
Kleinberg and Leighton [2003]		✓	$\widetilde{\mathcal{O}}(\sqrt{T})$
Javanmard and Nazerzadeh [2019]	✓		$\widetilde{\mathcal{O}}(s\sqrt{T})$
Shah et al. [2019]	✓ (log-linear model)	✓	$\widetilde{\mathcal{O}}(\sqrt{T}d^{11/4})$
Golrezaei et al. [2020]	✓	✓	$\widetilde{\mathcal{O}}(dT^{2/3})$ (changed benchmark)
Luo et al. [2021]	✓	✓	$\mathcal{O}_d(T^{2/3} + \lVert\widehat{\theta} - \theta_0\rVert_1 T)$
Our work	✓ (linear model)	✓	$\widetilde{\mathcal{O}}((Td)^{\frac{2m+1}{4m-1}})$

Table 1: Comparison with related works.

within the pricing algorithm, and we provide explicit rate on both the estimation error and the regret. Moreover, compared to several fully non-parametric dynamic pricing literatures, such as Mao et al. [2018] and Chen and Gallego [2020], our algorithm scales more nicely with dimension d, and can easily be generalized to a high-dimensional setting. Our algorithm is also easy to implement compared to some bandit-based algorithms that need dividing the feature space into bins.

Second, we generalize our results to the regime where the product features $\{\mathbf{x}_t\}_{t\geqslant 1}$ are weakly dependent instead of independent, which is more likely in practice. For example, for many products (such as softwares, electric products, etc.), the features of the products evolve over time and definitely inherit some past information. In other situations, the products for sale might have some common time-dependent factors shared by all products in the same industry (such as weather condition, population composition, etc.). This setting with weakly-dependent features can also be found in literatures such as Chen et al. [2022], where the authors study an offline pricing problem with parametric models and dependent covariates.

Last but not least, we establish non-asymptotic results on the ℓ_∞ error bound of the nonparametric kernel density and regression estimation, which are potentially useful in other related study as well. As mentioned in the related literatures, most results on non-parametric kernel regression estimation are established under the asymptotic settings. Meanwhile, we believe that non-asymptotic results are necessary to achieve a finite-sample regret upper bound in the pricing problem. Please refer to Appendix C.2 for related lemmas.

1.3 Notation

Throughout this work, we use $[n]$ to denote $\{1, 2, \cdots, n\}$. For any vector $\mathbf{x} \in \mathbb{R}^n$ and $q \geqslant 0$, we use $\|\mathbf{x}\|_q$ to represent the vector ℓ_q norm, i.e. $\|\mathbf{x}\|_q = (\sum_{i=1}^n |x_i|^q)^{1/q}$. In addition, we let $\nabla_{\mathbf{x}} L(\cdot), \nabla_{\mathbf{x}}^2 L(\cdot)$ be the gradient vector and Hessian matrix of loss function $L(\cdot)$ with respect to $\mathbf{x}$. For any given matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, we use $\|\cdot\|$ to denote the spectral norm of $\mathbf{X}$ and we write $\mathbf{X} \succcurlyeq 0$ or $\mathbf{X} \preccurlyeq 0$ if $\mathbf{X}$ or $-\mathbf{X}$ is semidefinite. For any event A, we let $\mathbb{I}_A$ be a indicator random variable which is equal to 1 if A is true and 0 otherwise. In addition, we use $\mathbb{C}^{(m)}$ with $m \in \mathbb{N}$ to denote the function class which contains all functions with m-th order continuous derivatives. For two positive sequences $\{a_n\}_{n \geqslant 1}$, $\{b_n\}_{n \geqslant 1}$, we write $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ if there exists a positive constant C such that $a_n \leqslant C \cdot b_n$ and we write $a_n = o(b_n)$ if $a_n / b_n \to 0$. In addition, we write $a_n = \Omega(b_n)$ or $a_n \gtrsim b_n$ if $a_n / b_n \geqslant c$ with some constant $c > 0$. We use $a_n = \Theta(b_n)$ if $a_n = \mathcal{O}(b_n)$ and $a_n = \Omega(b_n)$. We use notations $\mathcal{O}_d(\cdot), \Omega_d(\cdot)$ and $\Theta_d(\cdot)$ to denote similar meanings as above while treating the variable d as fixed. Moreover, we let $\widetilde{\mathcal{O}}(\cdot), \widetilde{\Omega}(\cdot), \widetilde{\Theta}(\cdot)$ represent the same meaning with $\mathcal{O}(\cdot), \Omega(\cdot)$ and $\Theta(\cdot)$ except for ignoring log factors.

1.4 Roadmap

The rest of this paper is organized as follows. We describe the problem in §2 and propose a solution in §3 where some heuristic arguments are offered for bounding the regret. In §4, we provide our theoretical results on the upper bounds of the regret and in §B, we discuss a lower bound result. Our algorithm is illustrated in §5 by intensive simulation experiments.

2 Problem Setting

We consider the pricing problem where a seller has a single product for sale at each time period $t = 1, 2, \cdots, T$. Here T is the total number of periods (i.e. length of horizon) and may be unknown to the seller. The market value of the product at time t is v_t and is unknown. We assume that the range of v_t is contained in a closed interval in $(0, B)$. In particular, we assume that $v_t \in [\delta_v, B - \delta_v]$ for some constant $\delta_v > 0$. At each period t, the seller posts a price p_t. If $p_t \leqslant v_t$, a sale occurs, and the seller collects a revenue of p_t; otherwise, no sale occurs and no revenue is obtained. Let y_t be the response variable that indicates whether a sale has occurred at period t. Then

$$y_t = \begin{cases} +1 & \text{if } v_t \geqslant p_t \, , \\ 0 & \text{if } v_t < p_t \, . \end{cases} \tag{1}$$

The goal of the seller is to design a pricing policy that maximizes the collected revenue.

In this paper, we further model the market value v_t as a linear function of the product's observable feature covariate $\mathbf{x}_t \in \mathbb{R}^d$. In particular, define $\widetilde{\mathbf{x}}_t = (\mathbf{x}_t^\top, 1)^\top$, where we assume $\{\mathbf{x}_t\}_{t \geqslant 1}$ are i.i.d. samples from an unknown distribution $\mathbb{P}_X$ supported on a bounded subset $\mathcal{X} \subseteq \mathbb{R}^d$. Assume that

$$v_t = \boldsymbol{\theta}_0^\top \widetilde{\mathbf{x}}_t + z_t, \tag{2}$$

where $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^\top, \alpha_0)^\top \in \mathbb{R}^{d+1}$ is an unknown parameter, and $\{z_t\}_{t \geqslant 1}$ is an i.i.d. sequence of idiosyncratic noise drawn from an **unknown** distribution with zero mean and bounded support $(-\delta_z, \delta_z)$. The cumulative distribution function of z_t is denoted by $F(\cdot)$. The above model implies that

$$y_t = \begin{cases} +1 & \text{with probability } 1 - F\left(p_t - \boldsymbol{\theta}_0^\top \widetilde{\mathbf{x}}_t\right) \, , \\ 0 & \text{with probability } F\left(p_t - \boldsymbol{\theta}_0^\top \widetilde{\mathbf{x}}_t\right) . \end{cases} \tag{3}$$

Remark 1. In fact, each $\mathbf{x}_t$ here can contain both product information and the buyer information, as long as this information is revealed to the seller.

Remark 2. The reason that we assume z_t has bounded support $[-\delta_z, \delta_z]$ is to ensure the market valuation $v_t \geqslant 0$, which is more reasonable in practice (Otherwise v_t has positive probability to be negative, since z_t is independent with the covariates $\mathbf{x}_t$). The truncated Gaussian distribution falls in such category. If the market allows v_t, p_t to be negative, then we can replace the boundness of z_t by any sub-Gaussian distributions.

In a non-dynamic setting, the model (3) is closely related to the single index model, or generalized linear (logistic regression) model with unknown link function [Ichimura, 1993, Fan et al., 1995, Carroll et al., 1997]. In their works, it's usually assumed that $p_t = 0$ and $\{(\widetilde{\mathbf{x}}_t)\}_{t\geqslant 1}$ are independent observations, and the goal is to estimate $\boldsymbol{\theta}_0$ and F. Meanwhile, we work on the dynamic setting where we need to optimize some revenue function by iteratively deciding p_t given previous observations based on dynamically learned $\boldsymbol{\theta}_0$ and F. These two problems are closely related but also decisively different.

We now state our objective in more details. Given observed features $\mathbf{x}_t$, the expected revenue at time t with a posted price p is

$$\mathrm{rev}_t(p, \boldsymbol{\theta}_0, F) := \mathbb{E} p \cdot \mathbb{1}(v_t \geqslant p) = p(1 - F(p - \boldsymbol{\theta}_0^\top \widetilde{\mathbf{x}}_t)). \tag{4}$$

The optimal posted price p_t^* for a product with attribute $\mathbf{x}_t$ is given by

$$p_t^* = \operatorname*{argmax}_{p\geqslant 0} p(1 - F(p - \boldsymbol{\theta}_0^\top \widetilde{\mathbf{x}}_t)), \tag{5}$$

which depends on unknown parameters and needs to be learned dynamically from the data. As in common practice, we evaluate the performance of any policy π that governs the rule of posted prices $\{p_t\}_{t\geqslant 1}$ by investigating the regret compared to the 'oracle pricing policy' that uses the knowledge of both $\boldsymbol{\theta}_0$ and $F(\cdot)$ and offers p_t^* according to (5) for any given t. In other words, we consider the problem of maximizing revenue as minimizing the following maximum regret

$$\mathrm{Regret}_\pi(T) \equiv \max_{\substack{\boldsymbol{\theta}_0 \in \Omega \\ \mathbb{P}_X \in \mathcal{Q}(\mathcal{X})}} \mathbb{E}\left[\sum_{t=1}^{T} \left(p_t^* \, \mathbb{1}(v_t \geqslant p_t^*) - p_t(\pi) \, \mathbb{1}(v_t \geqslant p_t(\pi))\right)\right], \tag{6}$$

where the expectation is taken with respect to the the idiosyncratic noise z_t and $\mathbf{x}_t$, and $p_t(\pi)$ denotes the price offered at time t by following policy π. Here $\mathcal{Q}(\mathcal{X})$ represents the set of probability distributions supported on a bounded set $\mathcal{X}$. Our goal is to choose a good strategy π such that the above total regret is small.

Apparently, learning $\boldsymbol{\theta}_0$ and $F(\cdot)$ over time gives the seller much more information to estimate the market value of a new product given it's feature covariates. On the other hand, the seller also wants to always give optimized price so as to maximize the expected revenue by (5). Therefore, it's necessary to have a good policy that strikes a balance between exploration (collecting data information for learning parameters) and exploitation (offering optimal pricing based on learned parameters).

Before proposing our algorithm, we first impose some regularity condition on F so that the optimization problem (5) is 'well-behaved'.

Assumption 1. *There exists a positive constant c_ϕ such that $\phi'(u) \geqslant c_\phi$ for all $u \in (-\delta_z, \delta_z)$, where $\phi(u) := u - \frac{1-F(u)}{F'(u)}$.*

Assumption 1 ensures that $\phi(\cdot)$ is strictly increasing, which implies a unique solution to (5). In fact, the first order condition of (5) yields

$$p_t^* = g(\boldsymbol{\theta}_0^\top \widetilde{\mathbf{x}}_t),$$

where $g(u) \triangleq u + \phi^{-1}(-u)$.

Remark 3. We only put some necessary assumptions on F in order to guarantee the existence of the unique optimal price p_t^* in (5), given observed $\widetilde{\mathbf{x}}_t$ and unknown but fixed $\boldsymbol{\theta}_0$. Comparing to the Assumption 2.1 in Javanmard and Nazerzadeh [2019], our Assumption 1 is weaker, since assumption that $1 - F(u)$ is log-concave is a special case of our assumption with $c_\phi \geqslant 1$.

3 Algorithm and Basic Regret Analysis

We first propose Algorithm 1 in §3.1 which describes our policy for minimizing the regret given in (6), and then provide the main idea for the regret analysis achieved by our Algorithm 1 in §3.2.

3.1 A Proposed Algorithm

In the following algorithm, we divide the time horizon into 'episodes' with increasing lengths. The first part of each episode is a short exploration phase where the offered prices are i.i.d. to collect the data and model parameters (i.e. $\widehat{\boldsymbol{\theta}}$, $\widehat{F}$) are then updated based on the collect data. The second part is an exploitation phase, where the optimal p_t is offered according to the current estimate of parameters and the new $\widetilde{\mathbf{x}}_t$. The details are stated in Algorithm 1.

Algorithm 1: Feature-based dynamic pricing with unknown noise distribution

```
Input: Upper bound of market value ($\{v_t\}_{t\geqslant 1}$): $B > 0$, minimum episode length: $\ell_0$, degree of smoothness: $m$.
Initialization: $p_1 = 0$, $\widehat{\boldsymbol{\theta}}_1 = 0$.
for each episode $k = 1, 2, \ldots,$ do
    Set length of the $k$-th episode $\ell_k = 2^{k-1}\ell_0$; Length of the exploration phase $a_k = \lceil (\ell_k d)^{\frac{2m+1}{4m-1}} \rceil$.
    Exploration Phase ($t \in I_k := \{\ell_k, \cdots, \ell_k + a_k - 1\}$):
        Offer price $p_t \sim \text{Unif}(0, B)$.
    Updating Estimates (at the end of the exploration phase with data $\{(\widetilde{\mathbf{x}}_t, y_t)\}_{t\in I_k}$):
        Update estimate of $\boldsymbol{\theta}_0$ by $\widehat{\boldsymbol{\theta}}_k = \widehat{\boldsymbol{\theta}}_k(\{(\widetilde{\mathbf{x}}_t, y_t)\}_{t\in I_k})$;
            $\widehat{\boldsymbol{\theta}}_k = \underset{\boldsymbol{\theta}}{\text{argmin}}\, L_k(\boldsymbol{\theta}) := \frac{1}{|I_k|}\sum_{t\in I_k}(By_t - \boldsymbol{\theta}^\top \widetilde{\mathbf{x}}_t)^2$    (7)
        Update estimates of $F$, $F'$ by $F_k(u, \widehat{\boldsymbol{\theta}}_k) = F_k(u; \widehat{\boldsymbol{\theta}}_k, \{(\widetilde{\mathbf{x}}_t, y_t, p_t)\}_{t\in I_k})$,
        $F_k^{(1)}(u, \widehat{\boldsymbol{\theta}}_k) = F_k^{(1)}(u, \widehat{\boldsymbol{\theta}}_k, \{(\widetilde{\mathbf{x}}_t, y_t, p_t)\}_{t\in I_k})$. The detailed formulas are given by (14) and (16).
        Update estimate of $\phi$ by $\widehat{\phi}_k(u) = u - \frac{1-\widehat{F}_k(u)}{\widehat{F}^{(1)}(u)}$ and estimate of $g$ by $\widehat{g}_k(u) = u + \widehat{\phi}_k^{-1}(-u)$.
    Exploitation Phase ($t \in I'_k := \{\ell_k + a_k, \cdots, \ell_{k+1} - 1\}$):
        Offer $p_t$ as
            $p_t = \min\{\max\{\widehat{g}_k(\widetilde{\mathbf{x}}_t^\top \widehat{\boldsymbol{\theta}}_k), 0\}, B\}.$    (8)
end for
```

Despite semiparametric model (3) with unknown link, by offering $p_t \sim \text{Unif}(0, B)$, By_t follows the linear model with regression $\widetilde{\mathbf{x}}_t^\top \boldsymbol{\theta}_0$ and this leads to the least-squares estimate (7). To see this, it follows that

$$\mathbb{E}[By_t \,|\, \widetilde{\mathbf{x}}_t] = B\mathbb{E}_{z_t}\mathbb{E}[y_t \,|\, \widetilde{\mathbf{x}}_t, z_t] = B\mathbb{E}_{z_t}\mathbb{E}[\mathbb{1}(p_t \leqslant \boldsymbol{\theta}_0^\top \widetilde{\mathbf{x}}_t + z_t) \,|\, \widetilde{\mathbf{x}}_t, z_t] = B\mathbb{E}\frac{\boldsymbol{\theta}_0^\top \widetilde{\mathbf{x}}_t + z_t}{B} = \widetilde{\mathbf{x}}_t^\top \boldsymbol{\theta}_0.$$

On the other hand, a uniform distribution for p_t is actually critical for the above property. Suppose that p_t is drawn from a c.d.f. $F_p(\cdot)$ and there is a transform f_1 of y_t that satisfies

$$\mathbb{E}f_1(y_t) = \mathbb{E}\tilde{\mathbf{x}}_t^\top \boldsymbol{\theta}_0 = \mathbb{E}v_t$$

for all $\mathbb{P}_X$, then according to (3), we have

$$\begin{aligned}\mathbb{E}v_t &= \mathbb{E}\mathbb{E}[f_1(y_t)\,|\,\tilde{\mathbf{x}}_t, z_t] = \mathbb{E}\mathbb{E}[f_1(\mathbb{1}(p_t \leqslant \tilde{\mathbf{x}}^\top \boldsymbol{\theta}_0 + z_t))\,|\,\tilde{\mathbf{x}}_t, z_t] \\ &= \mathbb{E}F_p(\tilde{\mathbf{x}}^\top \boldsymbol{\theta}_0 + z_t) f_1(1) + \mathbb{E}(1 - F_p(\tilde{\mathbf{x}}^\top \boldsymbol{\theta}_0 + z_t)) f_1(0) \\ &= f_1(0) + (f_1(1) - f_1(0))\mathbb{E}F_p(v_t).\end{aligned}$$

Since the above equation holds for all $\mathbb{P}_X \in \mathcal{Q}(X)$, it can only be the case that F_p is linear within the region $[0, B]$, which implies that p_t should follow a uniform distribution.

Remark 4. In Algorithm 1, the interval $[0, B]$ can be replaced with any interval that covers the range of the market value v_t. In practice, we can shrink the sampling interval at each exploration phase according to the feedback information observed in the past.

Remark 5. If z_t follows distributions with unbounded support and sub-Gaussian tails, in Algorithm 1, we only need to replace B by $B_k = C\sqrt{\log |I_k|}$ such that v_t falls in $(-B_k, B_k)$ with high probability. We then offer $p_t \sim$ Unif$(-B_k, B_k)$. Conditional on $v_t \in (-B_k, B_k)$, $B_k(2y_t - 1)$ serves as an unbiased estimator for $\tilde{\mathbf{x}}_t^\top \boldsymbol{\theta}_0$. Thus, all the following theoretical results work.

3.2 Main Idea for Regret Analysis

The main idea behind our regret analysis is a balance between exploration and exploitation. This idea is shown in the following heuristic arguments. For simplicity, we assume for now that there is only one episode, and that the total length of time (horizon) ℓ is known and d is bounded.

First, denote ℓ_1 as the length of the exploration phase. During this phase, the regret R_1 at each time is bounded by a constant due to bounded distribution $F(\cdot)$ that entails bounded p_t^* in (5). Therefore, the total regret in this phase is

$$R_1 = \mathcal{O}(\ell_1). \tag{9}$$

For the second phase, the expected regret can be controlled by the estimation error of both $\boldsymbol{\theta}$ and g (which is a functional of F as mentioned in (8)). In fact, let the regret at each time point t be

$$R_t := p_t^* \mathbb{I}_{(v_t \geqslant p_t^*)} - p_t \mathbb{I}_{(v_t \geqslant p_t)}.$$

Then the conditional expectation of regret at time t given previous information and $\widetilde{\mathbf{x}}_t$ is

$$\begin{aligned}\mathbb{E}[R_t \,|\, \bar{\mathcal{H}}_{t-1}] &= \mathbb{E}[p_t^* \mathbb{I}_{(v_t \geqslant p_t^*)} - p_t \mathbb{I}_{(v_t \geqslant p_t)} \,|\, \bar{\mathcal{H}}_{t-1}] \\ &= p_t^*(1 - F(p_t^* - \widetilde{\mathbf{x}}_t^\top \boldsymbol{\theta}_0)) - p_t(1 - F(p_t - \widetilde{\mathbf{x}}_t^\top \boldsymbol{\theta}_0)) \\ &= \mathrm{rev}_t(p_t^*, \boldsymbol{\theta}_0, F) - \mathrm{rev}_t(p_t, \boldsymbol{\theta}_0, F) \qquad (10)\end{aligned}$$

Here $\bar{\mathcal{H}}_t = \sigma(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{t+1}; z_1, \cdots, z_t)$. On the other hand, under mild conditions, the above difference in revenue can further be upper bounded by an order of $(p_t - p_t^*)^2$ using Taylor expansion. Therefore, we have

$$\begin{aligned}\mathbb{E}[R_t | \bar{\mathcal{H}}_{t-1}] \lesssim (p_t - p_t^*)^2 &= (\widehat{g}(\widehat{\boldsymbol{\theta}}^\top \widetilde{\mathbf{x}}_t) - g(\boldsymbol{\theta}_0^\top \widetilde{\mathbf{x}}_t))^2 \\ &\leqslant 2(\widehat{g}(\widehat{\boldsymbol{\theta}}^\top \widetilde{\mathbf{x}}_t) - g(\widehat{\boldsymbol{\theta}}^\top \widetilde{\mathbf{x}}_t))^2 + 2(g(\widehat{\boldsymbol{\theta}}^\top \widetilde{\mathbf{x}}_t) - g(\boldsymbol{\theta}_0^\top \widetilde{\mathbf{x}}_t))^2 \qquad (11) \\ &:= \mathbf{J_1} + \mathbf{J_2}.\end{aligned}$$

In fact, $\mathbf{J_2}$ is upper bounded by $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2^2$ (given the Lipschitz property of g according to Assumption 1 and suitable conditions over $\mathbb{P}_X$). By solving (7), we prove that the squared ℓ_2 error is of order $\mathcal{O}(\ell_1^{-1})$, which is the order of $\mathbf{J_2}$. The term $\mathbf{J_1}$ is upper bounded by $\|\widehat{g} - g\|_\infty^2$, and is further bounded by $\max\{\|\widehat{F} - F\|_\infty^2, \|\widehat{F}' - F'\|_\infty^2\}$. Note that by (1), $F(\cdot)$ is the non-parametric function of $1 - Y_t$ given $w_t = p_t - \widetilde{\mathbf{x}}_t^\top \boldsymbol{\theta}_0$, in which p_t is the observed price given in the exploration phase. Since $\boldsymbol{\theta}_0$ is estimated at a faster rate, we can assume that w_t is observable given a proper estimator of $\boldsymbol{\theta}_0$. Therefore, the error rate is dominated by estimating $F'(\cdot)$. Assuming F has an m-th continuous derivative, we construct $\widehat{g}$ using the kernel estimator with a m-th order kernel, and prove that $\max\{\|\widehat{F} - F\|_\infty, \|\widehat{F}' - F'\|_\infty\} \lesssim \mathcal{O}(\ell_1^{-(m-1)/(2m+1)})$ in which a logarithmic order is ignored for simplicity of presentation. Therefore, the total regret during the exploitation phase can be upper bounded by

$$R_2 \lesssim \ell \cdot \ell_1^{-2(m-1)/(2m+1)}. \qquad (12)$$

Combining (9) and (12), we know that by choosing ℓ_1 of the order of $\ell^{(2m+1)/(4m-1)}$, we balance the regret of both exploration and exploitation phase, and the total regret during the episode is given by

$$R_1 + R_2 = \mathcal{O}(\ell^{(2m+1)/(4m-1)}).$$

For a second order kernel, the above regret is of order $\mathcal{O}(\ell^{5/7})$. For a relatively large m, the regret is close to $\mathcal{O}(\ell^{1/2})$, which is actually proven to be the lower bound for a wider class of problems.

4 Regret Results on Proposed Policy

In this section, we divide our results into three parts. In §4.1, we consider the setting with independent covariates and finite differentiable noise distributions. In §4.2, we further extend our results in §4.1 to the setting with correlated features. Finally we extend the aforementioned results to the regime with infinitely differentiable noise distributions i.e. $m = \infty$ in §4.3.

4.1 Result under Independence Settings

The main result of this section is Theorem 1. To obtain this results, we first state some technical conditions and technical lemmas, which demonstrate the accuracy of statistical learning in each episode. These lemmas provide insights how statistical accuracy influences on the regret of our policy and have interests of their own rights.

Assume that $\|\boldsymbol{\theta}_0\| \leqslant R_\Theta$ for some constant $R_\Theta > 0$. We also define $R_{\mathcal{X}} := \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_2$. Before stating our main results, we first make the following assumptions on $\mathbf{x}_t$.

Assumption 2. *There exist positive constants $c_{\min}$ and $c_{\max}$, such that the covariance matrix $\boldsymbol{\Sigma}$ given by $\boldsymbol{\Sigma} = \mathbb{E}[\widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^\top]$ satisfies $c_{\min}\mathbb{I} \preccurlyeq \boldsymbol{\Sigma} \preccurlyeq c_{\max}\mathbb{I}$, where $\widetilde{\mathbf{x}}_t = (\mathbf{x}_t^\top, 1)^\top$*

As we observe from $\mathbf{J}_1, \mathbf{J}_2$ given in (11), bounding the regret in the exploitation phase needs to estimate both parameter $\boldsymbol{\theta}_0$ and function $g(\cdot)$. In the following, we first present an upper bound of estimating $\boldsymbol{\theta}_0$ at the end of the exploration phase within each episode in the following Lemma 1. Recall $|I_k|$ is the length of the k-th exploration phase.

Lemma 1. *Under Assumption 2, there exist positive constants c_0 and c_1 depending only on absolute constants given in assumptions such that for any episode k, as long as $|I_k| \geqslant c_0(d+1)$, with probability at least $1 - 2e^{-c_1 c_{\min}^2 |I_k|/16} - 2/|I_k|$,*

$$\|\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0\|_2 \leqslant \frac{8 \max\{R_{\mathcal{X}}, 1\}(R_{\mathcal{X}} R_\Theta + B)}{c_{\min}} \sqrt{\frac{(d+1)\log|I_k|}{|I_k|}}. \tag{13}$$

Let $\Theta_k := B(\boldsymbol{\theta}_0, R_k)$, where R_k is the right hand side of (13). We conclude from Lemma 1 that with high probability, R_k is of order at most $\sqrt{d\log|I_k|/|I_k|}$, and we can achieve similar upper bounds for $\mathbf{J}_2$ for any episode k.

Next, we proceed to construct the estimator $\widehat{g}_k$ in each episode and bound its distance to g. Notice that $g(u) = u + \phi^{-1}(-u)$, and $\phi(u) = u - \frac{1-F(u)}{F'(u)}$. Thus, a natural way to construct $\widehat{g}_k$ is from an estimate of F and F', as mentioned in our algorithm. Moreover, the uniform error bounds of our estimators $\widehat{F}_k$ and $\widehat{F}_k^{(1)}$ guarantee a uniform error bound of $\widehat{g}_k$.

We use the kernel regression method and $\widehat{\boldsymbol{\theta}}_k$ obtained above to construct $\widehat{F}_k$ and $\widehat{F}_k^{(1)}$. Recall that by (3), we have $E(y_t|w_t(\boldsymbol{\theta}_0)) = 1 - F(w_t(\boldsymbol{\theta}_0))$ where $w_t(\boldsymbol{\theta}) := p_t - \widetilde{\mathbf{x}}_t^\top \boldsymbol{\theta}$. Recall p_t is the observed price offered in the k-th exploration phase. Thus, given $\widehat{\boldsymbol{\theta}}_k$, $F(\cdot)$ can be estimated by using the Nadaraya-Watson kernel regression estimator and $F'(\cdot)$ can be estimated by the derivative of the estimator. Specifically, we define

$$\widehat{F}_k(u, \boldsymbol{\theta}) = 1 - \widehat{r}_k(u, \boldsymbol{\theta}) = 1 - \frac{h_k(u, \boldsymbol{\theta})}{f_k(u, \boldsymbol{\theta})}, \tag{14}$$

and $\widehat{F}_k(u) = \widehat{F}_k(u, \widehat{\boldsymbol{\theta}}_k)$, where

$$h_k(u, \boldsymbol{\theta}) = \frac{1}{|I_k| b_k} \sum_{t \in I_k} K(\frac{w_t(\boldsymbol{\theta}) - u}{b_k}) Y_t, \qquad f_k(u, \boldsymbol{\theta}) = \frac{1}{|I_k| b_k} \sum_{t \in I_k} K(\frac{w_t(\boldsymbol{\theta}) - u}{b_k}), \tag{15}$$

for a chosen m-th order kernel K and a suitable bandwidth b_k. Now, we estimate the derivative $F'(\cdot)$ by taking the derivative of the estimator. That is, $\widehat{F}_k^{(1)}(u) = \widehat{F}_k^{(1)}(u, \widehat{\boldsymbol{\theta}}_k)$ where

$$\widehat{F}_k^{(1)}(u, \boldsymbol{\theta}) = -\widehat{r}_k^{(1)}(u, \boldsymbol{\theta}) = -\frac{h_k^{(1)}(u, \boldsymbol{\theta}) f_k(u, \boldsymbol{\theta}) - h_k(u, \boldsymbol{\theta}) f_k^{(1)}(u, \boldsymbol{\theta})}{f_k^2(u, \boldsymbol{\theta})}, \tag{16}$$

$$h_k^{(1)}(u, \boldsymbol{\theta}) = \frac{-1}{|I_k| b_k^2} \sum_{t \in I_k} K'(\frac{w_t(\boldsymbol{\theta}) - u}{b_k}) Y_t, \qquad f_k^{(1)}(u, \boldsymbol{\theta}) = \frac{-1}{|I_k| b_k^2} \sum_{t \in I_k} K'(\frac{w_t(\boldsymbol{\theta}) - u}{b_k}). \tag{17}$$

Recall we mention in §2 that $(-\delta_z, \delta_z)$ is the support of noise z_t. In addition, we also mentions that T denotes the length of time horizon which is unknown. In the following, we will state other necessary assumptions to derive the regret upper bound:

Assumption 3. *The density of $w_t(\boldsymbol{\theta})$ (denoted as $f_{\boldsymbol{\theta}}$) satisfies the following:*

- *(Smoothness) There exists an integer $m \geqslant 2$ and a constant l_f such that for all $\boldsymbol{\theta} \in \Theta_0 := \big\{\boldsymbol{\theta} \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leqslant C_{\boldsymbol{\theta}} T^{-\frac{2m+1}{4(4m-1)}} d^{\frac{m-1}{4m-1}} \sqrt{\log T + 2\log d}\big\}$, $f_{\boldsymbol{\theta}}(u) \in \mathbb{C}^{(m)}$, and $f_{\boldsymbol{\theta}}^{(m)}$ is l_f-Lipschitz on $I := [-\delta_z, \delta_z]$.*

- *(Boundedness) There exists a constant $\bar{f} > 0$ such that $\forall u \in \mathbb{R}$ and $\boldsymbol{\theta} \in \Theta_0$, $\max\{|f_{\boldsymbol{\theta}}(u)|, |f'_{\boldsymbol{\theta}}(u)|\} \leqslant \bar{f}$. In addition, there exists a universal constant $c > 0$ such that $f_{\boldsymbol{\theta}}(u) \geqslant c$ for all $u \in I$, $\boldsymbol{\theta} \in \Theta_0$.*

Remark 6. We provide some examples for Assumption 3. For any covariate $\mathbf{x} \in \mathbb{R}^d$, as long as there exists an entry of it that follows a continuous distribution in $\mathbb{C}^{(m)}$, $m \geqslant 1$, such as Beta-distribution or truncated Gaussian distribution, we can ensure the density of $w(\boldsymbol{\theta}) = p_t - \tilde{\mathbf{x}}_t^\top \boldsymbol{\theta}$ satisfies both the smoothness and boundedness conditions in Assumption 3.

Assumption 4. *$r_{\boldsymbol{\theta}}(u) := \mathbb{E}[y_t \mid w_t(\boldsymbol{\theta}) = u]$ satisfies the following:*

- *(Smoothness) $h_{\boldsymbol{\theta}}(u) = f_{\boldsymbol{\theta}}(u) r_{\boldsymbol{\theta}}(u) \in \mathbb{C}^{(m)}$; $h_{\boldsymbol{\theta}}^{(m)}$ is l_f-Lipschitz on I for all $\boldsymbol{\theta} \in \Theta_0$. Here m and l_f are defined in Assumption 3.*

- *(Lipschitz) There exists a constant l_r such that $r_{\boldsymbol{\theta}_0} = 1 - F$ is l_r-Lipschitz, and for any $\epsilon > 0$, $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2 \leqslant \epsilon, u \in I} |r'_{\boldsymbol{\theta}}(u) - r'_{\boldsymbol{\theta}_0}(u)| \leqslant l_r \epsilon$.*

Assumption 5. *The kernel K satisfies the following:*

- *(Order-m kernel) $\int K(s)\mathrm{d}s = 1$, $\int s^j K(s)\mathrm{d}s = 0$ for $j \in \{1, \cdots, m-1\}$, and that $\int |s^m K(s)|\mathrm{d}s < +\infty$. Here m is the same as in Assumption 3.*

- *(Lipschitz) Both $K(s)$ and $K'(s)$ are l_K-Lipschitz continuous with bounded support.*

The Assumptions 3-5 are quite standard assumptions in non-parametric statistics; see Fan and Gijbels [1996], Tsybakov [2008] for more details. Given these assumptions, we will prove that with high probability, the estimators $\widehat{F}_k(u, \boldsymbol{\theta})$ and $\widehat{F}_k^{(1)}(u, \boldsymbol{\theta})$ are sufficiently close to $F(u)$ and $F'(u)$ respectively given any $\boldsymbol{\theta} \in \Theta_0$ for every sufficiently large k. Specifically, we obtain the desired error bound for $\widehat{F}_k(u) = \widehat{F}_k(u, \widehat{\boldsymbol{\theta}}_k)$ and $\widehat{F}_k^{(1)}(u) = \widehat{F}_k^{(1)}(u, \widehat{\boldsymbol{\theta}}_k)$.

Remark 7. Assumptions 3 and 4 can be relaxed in terms of the smoothness requirements: For all $m \geqslant 3$, we only need $f_{\boldsymbol{\theta}}(u), h_{\boldsymbol{\theta}}(u) \in \mathbb{C}^{(m-1)}$, and that $f_{\boldsymbol{\theta}}^{(m-1)}(u), h_{\boldsymbol{\theta}}^{(m-1)}(u)$ are ℓ-Lipschitz for some constant ℓ. For $m = 2$, we only need $f_{\boldsymbol{\theta}}(u), h_{\boldsymbol{\theta}}(u) \in \mathbb{C}^{(1)}$, and that the second order derivatives of $f_{\boldsymbol{\theta}}(u), h_{\boldsymbol{\theta}}(u)$ exist and are bounded. One is able to see assuming functions in $\mathbb{C}^{(m)}$ is a sufficient condition for the aforementioned conditions to hold, for the simplicity of our notations here, we keep the original assumptions.

Remark 8. If we only assume $F(\cdot)$ is ℓ-Lipschitz continuous (i.e. it may not be differentiable), we also provide an alternative algorithm in §H which achieves a regret upper bound $\widetilde{\mathcal{O}}(T^{3/4})$.

Remark 9. One is also able to estimate $F(u), F'(u)$ with the local polynomial estimator (see e.g. Fan and Gijbels [1996]). In this case, the assumptions can be weaken further. Specifically, the local polynomial estimators for F and F' enjoy all the theoretical guarantees given only the second part of Assumptions 3 and 5 instead of both Assumptions 3 and 5. For example, Lipschitz continuous density functions on $[-\delta_z, \delta_z]$ satisfy Assumption 4.2. The proof is very similar. For simplicity, we only focus on studying kernel regression in this paper.

Lemma 2. *Under Assumptions 3, 4 and 5, there exist constants $B_{x,K}$, $B'_{x,K}$ and $C_{x,K}$ (depending only the absolute constants within the assumptions) such that as long as*

$$T \geqslant B_{x,K}(\log T + 2\log d)^{\frac{4m-1}{m}} d^{\frac{2m-1}{m}},$$

we have for any $k \geqslant \lfloor (\log(\sqrt{T} + \ell_0) - \log \ell_0)/\log 2 \rfloor + 2$ and $\delta \in [4\exp(-B'_{x,K}|I_k|^{\frac{2m}{2m+1}}/\log|I_k|), \frac{1}{2}]$, with probability at least $1 - 2\delta$,

$$\sup_{u \in I, \boldsymbol{\theta} \in \Theta_k} |\widehat{F}_k(u, \boldsymbol{\theta}) - F(u)| \leqslant C_{x,K}|I_K|^{-\frac{m}{2m+1}}\sqrt{\log|I_K|}(\sqrt{d} + \sqrt{\log\frac{1}{\delta}}). \tag{18}$$

Here $I = [-\delta_z, \delta_z]$ and we choose the bandwidth $b_k = |I_k|^{-\frac{1}{2m+1}}$.

Lemma 3. *Under the same conditions as Lemma 2, with probability at least $1 - 4\delta$, we have*

$$\sup_{u \in I, \boldsymbol{\theta} \in \Theta_k} |\widehat{F}_k^{(1)}(u, \boldsymbol{\theta}) - F'(u)| \leqslant \widetilde{C}_{x,K} |I_K|^{-\frac{m-1}{2m+1}} \sqrt{\log |I_K|} (\sqrt{d} + \sqrt{\log \frac{1}{\delta}}). \tag{19}$$

We next develop a uniform upper bound for term $\mathbf{J}_1$ given in (11) for the k-th episode in Lemma 4 below.

Lemma 4. *Reinstating the notations and conditions in Lemma 2, with probability at least $1 - 6\delta$, we have*

$$\sup_{u \in [\delta_z, B - \delta_z]} |\widehat{g}_k(u) - g(u)| \leqslant \bar{C}_{x,K} |I_K|^{-\frac{m-1}{2m+1}} \sqrt{\log |I_K|} (\sqrt{d} + \sqrt{\log \frac{1}{\delta}}).$$

Remark 10. In Algorithm 1 we define $\widehat{g}_k(u) = u + \widehat{\phi}_k^{-1}(-u)$ with $u \in [\delta_z, B - \delta_z]$. Thus, computing $\widehat{g}_k(u)$ involves obtaining the inverse of $\widehat{\phi}_k$, which is not necessarily monotonic. Nevertheless, it's not difficult to define or compute $\widehat{\phi}_k^{-1}$. In fact, we'll show in the proof of Lemma 4 that $\widehat{\phi}_k$ is very 'close' to ϕ in some main interval of interest, which contains $[\phi^{-1}(\delta_z - B), \phi^{-1}(-\delta_z)]$ and depends only on F. (Recall in Assumption 1 that ϕ' is bounded below from 0, so ϕ is strictly increasing). Thus, for any $u \in [\delta_z, B - \delta_z]$, the above fact will guarantee the existence of $\widehat{\phi}_k^{-1}(-u)$ as some x within the interval such that $\widehat{\phi}_k(x) = -u$.

Combining the above lemmas, which give us upper bounds for terms $\mathbf{J}_1, \mathbf{J}_2$ in every episode, we have the following Theorem 1, which provides an upper bound for the regret.

Theorem 1. *Under Assumptions 1, 3, 4 and 5, there exist constants $\bar{B}_{x,K}$, $\bar{B}'_{x,K}$ and $C^*_{x,K}$ (depending only on the absolute constants within the assumptions) such that for all T satisfying*

$$T \geqslant \max\{\bar{B}_{x,K}(\log T + 2\log d)^{\frac{4m-1}{m-1}} d^{\frac{2m+1}{m-1}}, 4d^{\frac{2m+1}{m-1}}\},$$

the regret of Algorithm 1 over time T is no more than $C^*_{x,K}(Td)^{\frac{2m+1}{4m-1}} \log T(1 + \log T/d)$.

Remark 11. We note that Golrezaei et al. [2020] shares a similar framework with ours, although with a different regret measure. Specifically, we use a more traditional notion of regret by setting the benchmark p_t^* from (5) with true $\boldsymbol{\theta}_0$ and $F(\cdot)$. In Golrezaei et al. [2020], the authors instead set the benchmark p_t^* so as to maximize the worst function in their function class $\mathcal{F}$, i.e.

$$p_t^* = \operatorname*{argmax}_{p \geqslant 0} \min_{F \in \mathcal{F}} p(1 - F(p - \boldsymbol{\theta}_0^\top \widetilde{\mathbf{x}}_t)).$$

Their optimal regret is of order $\widetilde{\mathcal{O}}_d(T^{2/3})$, while ours is $\widetilde{\mathcal{O}}_d(T^{\frac{2m+1}{4m-1}})$, which is closer to $\mathcal{O}_d(T^{1/2})$ when m is sufficiently large. Intuitively, a benchmark being the price maximizing the worst function is too conservative when their ambiguity function class is very large and the market noises are only sampled from a fixed distribution function in that function class, which is true in our semi-parametric setting.

On the other hand, Golrezaei et al. [2019] also work on similar but simpler settings, where they assume having unknown demanding curves but observable valuations instead of censored responses. By contrast, we work on a more common setting where the actual market values of products are unknown.

Remark 12. Both Algorithm 1 and Theorem 1 depend on the smoothness class of the function $F(\cdot)$. A popular choice in nonparametric curve estimation literature is $m = 2$, as other choices do not improve much for practical sample sizes. Nevertheless, we provide two ways to choose m that addresses a referee's query.

- **Estimate m using cross-validation.** Specifically, we pick some relatively small m during the first episode. At each episode $k \geqslant 2$, before entering the exploration phase, we update the estimate of m using cross-validation [Hall and Racine, 2015] with the data gathered from the previous exploration phase. Then, we proceed with the main algorithm with this updated estimate until the next episode. For more details of the cross-validation procedure and the combined algorithm, see Section I.

- **Pick a constant pessimistic estimation of m.** In fact, we can directly fix a relatively small m (e.g. $m = 2$ or $m = 4$). In many cases, the performance of the algorithm ($\widetilde{\mathcal{O}}((Td)^{5/7})$ and $\widetilde{\mathcal{O}}((Td)^{3/5})$) will not be significantly different from where m is known (at least $\Omega((Td)^{1/2})$).

The above two ways can be applied to all settings in this paper as long as F is only required to be smooth to a finite degree.

4.2 Results under the setting with strong-mixing features

As mentioned in the introduction, we believe that in many situations, the dependence of features over time is inevitable. Thus, in this section, we generalize our results to the case where $\mathbf{x}_t$ can be dependent. For this purpose, we first impose the strong-mixing condition which measure the dependence between covariates over time.

Definition 1. *[β-mixing] For a sequence of random vectors $\mathbf{x}_t \in \mathbb{R}^{d\times 1}$ on a probability space $(\Omega, \mathcal{X}, \mathbb{P})$, define β-mixing coefficient*

$$\beta_k = \sup_{l \geqslant 0} \beta(\sigma(\mathbf{x}_t, t \leqslant l), \sigma(\mathbf{x}_t, t \geqslant l+k))$$

in which

$$\beta(\mathcal{A}, \mathcal{B}) = \frac{1}{2} \sup \Big\{ \sum_{i \in I} \sum_{j \in J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)| \Big\},$$

the maximum being taken over all finite partitions $(A_i)_{i\in I}$ and $(B_i)_{i\in J}$ of Ω with elements in $\mathcal{A}$ and $\mathcal{B}$.

The following assumption ensures that $\{\mathbf{x}_t\}_{t\geqslant 1}$ are not too strongly dependent. Combining with other assumptions, we ensure that the empirical covariance matrix $\frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^{\top}$ concentrate around the population version, which is necessary in deriving the regret in every episode.

Assumption 6. *The sequence $\mathbf{x}_t, t \geqslant 0$ are strictly stationary time series and follow β-mixing condition, in a sense we assume that $\beta_k \leqslant e^{-ck}$ holds with some constant c.*

In order to derive the final regret upper bound under the stong-mixing setting, we also need an additional technical assumption stated below:

Assumption 7. *Let $r_{\boldsymbol{\theta}}(u_i, u_j) := \mathbb{E}[y_i y_j \mid w_j(\boldsymbol{\theta}) = u_j, w_i(\boldsymbol{\theta}) = u_i]$, $j > i \geqslant 0$, $r_{\boldsymbol{\theta}}(u_j) := \mathbb{E}[y_j \mid w_j(\boldsymbol{\theta}) = u_j], j \geqslant 0$ be the joint regression function and marginal regression function. In addition, we also set $f_{\boldsymbol{\theta}}(u_i, u_j)$, $j > i \geqslant 0$, $f_{\boldsymbol{\theta}}(u_i), i \geqslant 0$ as the joint density of $w_i(\boldsymbol{\theta})$ and $w_j(\boldsymbol{\theta})$ and marginal density of $w_i(\boldsymbol{\theta})$ respectively. Then we define $g_{1,\boldsymbol{\theta}}(u_i, u_j) := r_{\boldsymbol{\theta}}(u_i, u_j) f_{\boldsymbol{\theta}}(u_i, u_j) - r_{\boldsymbol{\theta}}(u_i) f_{\boldsymbol{\theta}}(u_i) r_{\boldsymbol{\theta}}(u_j) f_{\boldsymbol{\theta}}(u_j)$*

and $g_{2,\boldsymbol{\theta}}(u_i, u_j) = f_{\boldsymbol{\theta}}(u_i, u_j) - f_{\boldsymbol{\theta}}(u_i) f_{\boldsymbol{\theta}}(u_j)$. We assume $g_{1,\boldsymbol{\theta}}(u_i, u_j)$ and $g_{2,\boldsymbol{\theta}}(u_i, u_j)$ follow l-Lipschitz continuous condition, in a sense that

$$|g_{q,\boldsymbol{\theta}}(u_i, u_j) - g_{q,\boldsymbol{\theta}}(u_i', u_j')| \leqslant l\sqrt{(u_i - u_i')^2 + (u_j - u_j')^2},\ q \in \{1, 2\}$$

holds for all (u_i, u_j), with $i, j \in [n]$ and $\boldsymbol{\theta} \in \Theta_0$.

When the covariates $\mathbf{x}_i, \mathbf{x}_j$ are independent, we have $g_{q,\boldsymbol{\theta}}(u_i, u_j) = 0, q \in \{1, 2\}$, for all (u_i, u_j). Under such a mild assumption, we obtain a uniform upper bound of $|g_{q,\boldsymbol{\theta}}(u_i, u_j)|$, which is dominated by the β-mixing constant $\beta_{j-i}^{1/3}$, for all $\boldsymbol{\theta} \in \Theta_0$ and (u_i, u_j) (see Appendix F.7). Thus, this assumption essentially guarantees that the joint regression and density functions of the features still stay close to the products of their marginal ones even if they are correlated.

Following similar analysis with §4.1, we reach the following theorem which gives a regret upper bound at similar rate with Theorem 1 under the strong-mixing feature setting.

Theorem 2. *Let Assumptions 1, 3, 4, 5, 6 and 7 hold. Then there exist constants $B_{mx,K}^*$ and $C_{mx,K}^*$ (depending only on the absolute constants within the assumptions) such that for all T satisfying*

$$T \geqslant \max\{B_{mx,K}^*(\log T + 2\log d)^{\frac{12m-3}{m-1}}[(d+1)\log(d+1)]^{\frac{4m-1}{m-1}}/d^2, d^{\frac{2m+1}{m-1}}\}$$

the regret of Algorithm 1 over time T is no more than $C_{mx,K}^*(Td)^{\frac{2m+1}{4m-1}}\log^4 T$.

4.3 Result on infinitely differentiable market noise distribution

In §4.1 and §4.2, we analyze the regret upper bounds when the noise distribution F has an m-th order continuous derivative, with any finite $m \geqslant 2$. The regret of our algorithm is of order $\widetilde{\mathcal{O}}((Td)^{\frac{2m+1}{4m-1}})$, which gets closer to $\widetilde{\mathcal{O}}(\sqrt{Td})$ as the degree of smoothness m goes to infinity. In fact, this is mainly due to inaccurate estimation of F and F' resulting from the bias of the kernel estimator. In this section, we deal with super smooth noise distributions [Fan, 1991], where F is infinitely differentiable. Under mild conditions, we're able to control the bias within $\mathcal{O}(1/|I_k|^{\frac{1}{2}})$ for each episode k by using extremely smooth kernels. As a reminder, here $|I_k|$ is the length of the k-th exploration phase. This leads to a $\widetilde{\mathcal{O}}_d(\sqrt{T})$ regret bound in our algorithm. In particular, we assume the following:

Assumption 8. *Define $\phi_{\boldsymbol{\theta}}$, $\xi_{\boldsymbol{\theta}}$, $\phi_{\boldsymbol{\theta}}^{(1)}$ and $\xi_{\boldsymbol{\theta}}^{(1)}$ as the Fourier transform of the function $f_{\boldsymbol{\theta}}$, $h_{\boldsymbol{\theta}}$, $f'_{\boldsymbol{\theta}}$ and $h'_{\boldsymbol{\theta}}$ respectively:*

$$\phi_{\boldsymbol{\theta}}(s) = \int_{-\infty}^{\infty} f_{\boldsymbol{\theta}}(x)e^{isx}\mathrm{d}x, \;\; \xi_{\boldsymbol{\theta}}(s) = \int_{-\infty}^{\infty} h_{\boldsymbol{\theta}}(x)e^{isx}\mathrm{d}x,$$
$$\phi_{\boldsymbol{\theta}}^{(1)}(s) = \int_{-\infty}^{\infty} f'_{\boldsymbol{\theta}}(x)e^{isx}\mathrm{d}x, \;\; \xi_{\boldsymbol{\theta}}^{(1)}(s) = \int_{-\infty}^{\infty} h'_{\boldsymbol{\theta}}(x)e^{isx}\mathrm{d}x,$$

and $h_{\boldsymbol{\theta}}(x) = f_{\boldsymbol{\theta}}(x)r_{\boldsymbol{\theta}}(x)$. There exist positive constant D_ϕ and d_ϕ and $\alpha > 0$ such that

$$\max\{|\phi_{\boldsymbol{\theta}}(s)|, |\xi_{\boldsymbol{\theta}}(s)|, |\phi_{\boldsymbol{\theta}}^{(1)}(s)|, |\xi_{\boldsymbol{\theta}}^{(1)}(s)|\} \leqslant D_\phi e^{-d_\phi |s|^\alpha}$$

for all $s \in \mathbb{R}$.

Remark 13. -This assumption is quite standard, and ensures that $f_{\boldsymbol{\theta}}(u)$, $F_{\boldsymbol{\theta}}(u) \in \mathbb{C}^\infty$. The class of functions are still infinite dimensional nonparametric functions. The class of supersmooth functions has been used in non-parametric density literature. In particular, it has been used in Fan [1991] for characterizing the difficulty of non-parametric deconvolution.

Under the Assumption of 8, for each episode k, we can successfully control the bias within $\mathcal{O}(1/\sqrt{|I_k|})$ via an infinite order kernel [McMurry and Politis, 2004, Berg and Politis, 2009]. In order to construct an infinite order kernel K, we simply let K be the Fourier inverse transform of some 'well-behaved' function. In particular, let

$$K(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \kappa(s)e^{-isx}\mathrm{d}s, \tag{20}$$

be the Fourier inversion of κ satisfying

$$\kappa(s) = \begin{cases} 1, & |s| \leqslant c_\kappa \\ g_\infty(|s|), & \text{otherwise.} \end{cases}$$

Here g_∞ is any continuous, square-integrable function that is bounded in absolute value by 1 and satisfies $g_\infty(|c_\kappa|) = 1$. This defines an infinity order kernel function [Fan and Gijbels, 1996].

By plugging the infinite order kernel K into our algorithm, we're able to obtain the following lemma:

Lemma 5. *Under Assumption 8, there exists a positive constant C_{inf} depending only on α, D_ϕ and d_ϕ such that for all kernel K satisfying (20), for each episode k, by choosing the bandwidth $b_k = c_\kappa (d_\phi / \log|I_k|)^{1/\alpha}$ in (15) and (17), we have*

$$\sup_{u\in I,\boldsymbol{\theta}\in\Theta_k} |\mathbb{E}[f_k(u,\boldsymbol{\theta})] - f_{\boldsymbol{\theta}}(u)| \leqslant \frac{C_{\text{inf}}}{\sqrt{|I_k|}}, \quad \sup_{u\in I,\boldsymbol{\theta}\in\Theta_k} |\mathbb{E}[h_k(u,\boldsymbol{\theta})] - h_{\boldsymbol{\theta}}(u)| \leqslant \frac{C_{\text{inf}}}{\sqrt{|I_k|}},$$
$$\sup_{u\in I,\boldsymbol{\theta}\in\Theta_k} |\mathbb{E}[f_k^{(1)}(u,\boldsymbol{\theta})] - f'_{\boldsymbol{\theta}}(u)| \leqslant \frac{C_{\text{inf}}}{\sqrt{|I_k|}}, \quad \sup_{u\in I,\boldsymbol{\theta}\in\Theta_k} |\mathbb{E}[h_k^{(1)}(u,\boldsymbol{\theta})] - h'_{\boldsymbol{\theta}}(u)| \leqslant \frac{C_{\text{inf}}}{\sqrt{|I_k|}}.$$

Following similar proof procedures of Theorems 1 and 2, Lemma 5 leads to the following theorem, which gives a regret upper bound of $\widetilde{\mathcal{O}}_d(\sqrt{T})$, achieving the same convergence rate with the parametric case up to logarithmic terms [Javanmard and Nazerzadeh, 2019].

Theorem 3. *Let Assumptions 1, 3, 4, 5, 6, 7 and 8 hold. Then there exist constants B^*_{inf} and C^*_{inf} (depending only on the absolute constants within the assumptions) such that by choosing $|I_k| = \lceil \sqrt{l_k d} \rceil$ instead in Algorithm 1, for all T satisfying*

$$T \geqslant B^*_{\text{inf}} d^2 (\log T + 2\log d)^{12+12/\alpha} \log^4(d+1),$$

the regret of the algorithm over time T is no more than $C^*_{\text{inf}} (Td)^{\frac{1}{2}} (\log T)^{\frac{3}{2}+\frac{3}{2\alpha}} [\log(d+1) + \log T/d]$.

Remark 14. Theorem 3 partly overturns the conjecture in Shah et al. [2019] that there is no policy can achieve an $\widetilde{\mathcal{O}}_d(\sqrt{T})$ regret under the setting where the market value is linear in the features as in (2). We provide a regime with super smooth market noise in which $\widetilde{\mathcal{O}}_d(\sqrt{T})$ regret upper bound is attainable by our policy.

Remark 15. In §B, we provide discussions on minimax lower bound, adversarial agents, inference for the demand, practical implementation, and extensions to the high dimensional setting.

5 Simulations

5.1 Justification of theoretical results

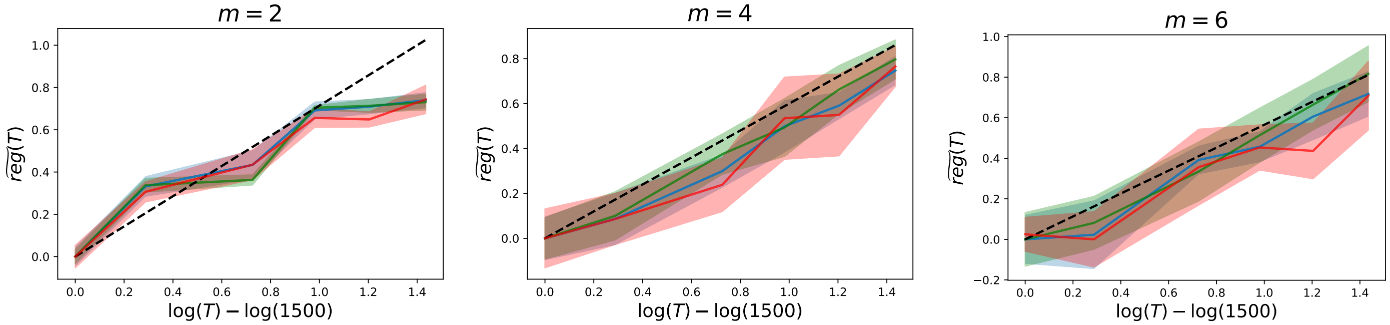
In this section, we illustrate the performance of our policy through large-scale simulations under various settings. Recall our model (2), where $\mathbf{x}_t \in \mathbb{R}^d$ and z_t follows distributions with bounded support and smooth c.d.f. Throughout this section, we let the dimension $d = 3$ and the coefficients $\alpha_0 = 3$, $\boldsymbol{\beta}_0 = \sqrt{2/3} \cdot \mathbf{1}_{3\times 1}$. For each value of smoothness degree $m \in \{2, 4, 6\}$, we fix a density function from $\mathbb{C}^{(m-1)}$ for all z_t (thus the c.d.f. F belongs to $\mathbb{C}^{(m)}$). Specifically, we set the p.d.f. of z_t as $f_m(x) \propto (1/4 - x^2)^{m/2} \cdot \mathbb{I}_{\{|x| \leqslant 1/2\}}$ for $m \in \{2, 4, 6\}$. Moreover, for each m, the covariates $\mathbf{x}_t \in \mathbb{R}^3$ are generated from a p.d.f. in $\mathbb{C}^{(m)}$ in the following ways:

- **i.i.d. $\mathbf{x}_t$ with independent entries:** Each coordinate of $\mathbf{x}_t$ is generated from density $f_m(x) \propto (2/3 - x^2)^{m+1} \cdot \mathbb{I}_{\{|x| \leqslant \sqrt{2/3}\}}$.

- **i.i.d. $\mathbf{x}_t$ with dependent entries:** $\mathbf{x}_t$ is generated from the density function $f_m(\mathbf{x}) \propto (1 - \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x})^{m+1}$. Here $\boldsymbol{\Sigma}$ is a positive definite matrix with (i, j)-th entry being equal to $0.2^{|i-j|}, 1 \leqslant i, j \leqslant 3$.

- **Strong mixing $\mathbf{x}_t$ with dependent entries:** We generate $\mathbf{x}_t$ from the VAR (vector autoregression) model, where $\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{x}_{t-2} + \boldsymbol{\xi}_t$. Here $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{3\times 3}$ with $\mathbf{A}_{i,j} = 0.4^{|i-j|+1}$, $\mathbf{B}_{i,j} = 0.1^{|i-j|+1}$, $i, j \in \{1, 2, 3\}$. In addition, $\{\boldsymbol{\xi}_t\}_{t\geqslant 1}$ are i.i.d. with density $f_m(\boldsymbol{\xi}) \propto (1 - \boldsymbol{\xi}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\xi})^{m+1}$ where the $\boldsymbol{\Sigma}$ is the same as the one given in **(ii)**.

When implementing our algorithm, we divide the time horizon into consecutive episodes by setting the length of the k-th episode as $\ell_k = 2^{k-1}\ell_0$ with $k \in \mathbb{N}^+$ and $\ell_0 = 200$. We further separate every episode into an exploration phase with length $|I_k| = \min\{(d\ell_k)^{(2m+1)/(4m-1)}, \ell_k\}$ depending on the values of m and d. The exploitation phase contains the rest of the time in that episode. In the exploration phase, we sample p_t from $\text{Unif}(0, B = 6)$, since $B = 6$ is a valid upper bound of v_t. In the exploitation phase, we set the kernels as follows: For any given $m \in \{2, 4, 6\}$ prefixed at the beginning of the algorithm, we choose the kernel function with m-th order. Here we choose the second, fourth, sixth-order kernel functions as $K_2(u) = 35/12(1 - u^2)^3 \cdot \mathbb{I}_{\{|u| \leqslant 1\}}$, $K_4(u) = 27/16(1 - 11/3u^2) \cdot K_2(u)$ and $K_6(u) = 297/128(1 - 26/3u^2 + 13u^4) \cdot K_2(u)$ respectively. In episode k, we set the bandwidth b_k as $3 \cdot |I_k|^{-\frac{1}{2m+1}}$ in (14) and (16) according to the settings in the theoretical analysis. In reality, one can also tune the bandwidth by using cross validation at the end of every exploration phase. Moreover, when calculating $p_t = \widehat{g}(\widetilde{\mathbf{x}}_t^\top \widehat{\boldsymbol{\theta}}_k) = \widetilde{\mathbf{x}}_t^\top \widehat{\boldsymbol{\theta}}_k + \widehat{\phi}_k^{-1}(-\widetilde{\mathbf{x}}_t^\top \widehat{\boldsymbol{\theta}}_k)$, we find $\widehat{\phi}_k^{-1}(-\widetilde{\mathbf{x}}_t^\top \widehat{\boldsymbol{\theta}}_k)$ as follows: First, we look for



(a) (b) (c)

Figure 1: Regret log-log plot in the setting with i.i.d. covariates with independent entries. The three subplots show the case $m \in [2, 4, 6]$ respectively. The x-axis is $\log(T) - \log(1500)$ for $T \in [1500, 2000, 3100, 4000, 5000, 6300]$, while the y-axis is $\widetilde{\text{reg}}(T) := \log(\text{reg}(T)) - 2 \log \log T - (\log(\text{reg}(1500)) - 2 \log \log 1500)$. The solid blue, green and red lines represent the mean $\widetilde{\text{reg}}(T)$ of the Algorithm 1 with unknown $g(\cdot)$ and $\boldsymbol{\theta}_0$, unknown $g(\cdot)$ but known $\boldsymbol{\theta}_0$, and known $g(\cdot)$ but unknown $\boldsymbol{\theta}_0$ respectively over 30 independent runs. The light color areas around those solid lines depict the standard error of our estimation of $\log(\text{reg}(T)) - 2 \log \log T$. The dashed black lines in $(a)-(c)$ represents the benchmark whose slopes are equal to $\frac{2m+1}{4m-1}$ with $m \in \{2, 4, 6\}$.

$x \in [-1, 1]$ such that $\widehat{\phi}_k(x) = -\widetilde{\mathbf{x}}_t^\top \widehat{\boldsymbol{\theta}}_k$ (The interval $[-1, 1]$ contains the true support of $\phi(x)$ [-0.5, 0.5], since in reality, we might only know a range of the true support). Then, we do a transformation of variable x to $x(y) = -2 \cdot \exp(y)/(1 + \exp(y)) + 1$ and solve y as the root of $\widehat{\phi}_k(x(y)) + \widetilde{\mathbf{x}}_t^\top \widehat{\boldsymbol{\theta}}_k = 0$ by using Newton's method starting at $y = 0$. Finally, we set $x = -2 \cdot \exp(y)/(1 + \exp(y)) + 1$ as $\widehat{\phi}_k^{-1}(-\widetilde{\mathbf{x}}_t^\top \widehat{\boldsymbol{\theta}}_k)$ and offer p_t according to the algorithm.

For any given $m \in \{2, 4, 6\}$, under the three covariate settings discussed above, we input m into the algorithm, select the corresponding kernel and repeat Algorithm 1 for 30 times until $T = 6300$. For each $T \in [1500, 2000, 3100, 4000, 5000, 6300]$, we record the cumulative regret reg(T). For the first two covariate settings, recall from Theorem 1 that the regret reg$(T) \lesssim T^{\frac{2m+1}{4m-1}} \log^2 T$. Thus, we plot $\widetilde{\text{reg}}(T)$ against $\log(T) - \log(1500)$ in Figure 1, 7, 8, where $\widetilde{\text{reg}}(T) := \log(\text{reg}(T)) - 2 \log \log T - (\log(\text{reg}(1500)) - 2 \log \log 1500)$;

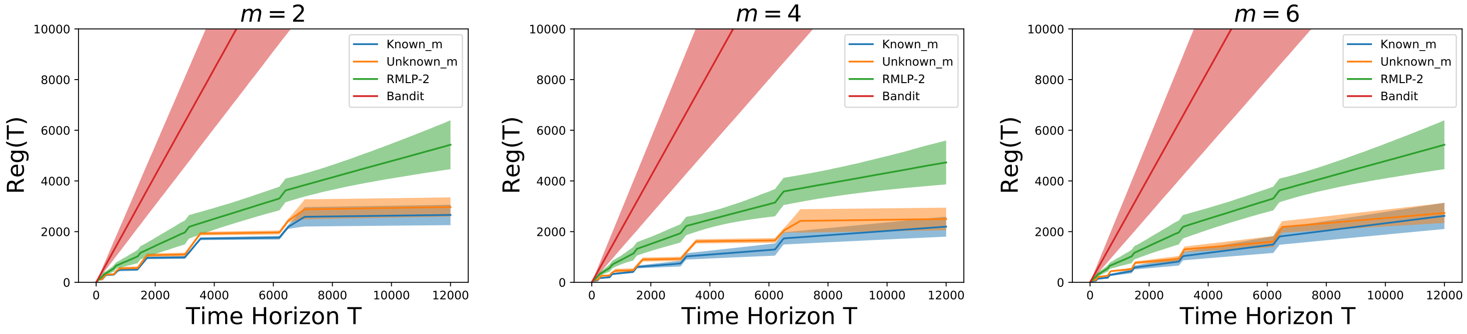
From Figures 1, 7, 8, we conclude that under all settings, the rates of the empirical regrets' increments produced by Algorithm 1 (as shown by the solid blue lines) do not exceed their theoretical counterparts given in Theorems 1 and 2 (as shown by the dashed black lines). In many cases, the growth rates of the empirical regrets are very close to those of the theoretical lines. This demonstrates the tightness of our theoretical results. Moreover, as all the solid lines have similar growth rates, we show that Algorithm 1 is robust to the estimation of $\boldsymbol{\theta}_0$ and $g(\cdot)$. This is further proved in Appendix G, where we directly plot reg(T) for all the settings discussed here. See Appendix G for more plots and discussions.

5.2 Comparison with other methods

In this subsection, we provide numerical studies which illustrate differences between our methods and two highly related prior arts ('RMLP-2' and 'Bandit') using both synthetic and real data. Here, 'RMLP-2' is the policy proposed in Javanmard and Nazerzadeh [2019] that solves the same problem as ours except that the noise distribution falls in a *parametric* function class. In addition, we denote the policy proposed in Kleinberg and Leighton [2003] as 'Bandit', which leverages a variant of UCB algorithm under non-parametric noise distribution that achieves $\mathcal{O}(\sqrt{T})$ regret *without* modeling covariate information.

We first use synthetic data to illustrate the efficiency of our method over 'RMLP-2' and 'Bandit'. For each smoothness degree $m = \{2, 4, 6\}$, we generate our data following the same way given in §5.1, except that we only generate the distribution of $\mathbf{x}_t$ according to the first option discussed in §5.1. We illustrate the performance of our method against those two prior arts in the following figures. Here we follow Algorithm 4 which uses a data-driven way to determine m before every episode. For RMLP-2, since there is no way the algorithm knows the true noise distribution, we instead assume the noise falls into a Gaussian distribution when executing the algorithm.

We see from the simulation results that the regret we achieved is much smaller than those two benchmarks. As for the comparison with RMLP-2, our method is robust to the mis-specification of the parametric function class since our algorithm can adapt to all functions in the non-parametric class. For the comparison with 'Bandit', we see that only using the non-parametric bandit algorithm without considering the contextual information (heterogeneity of product) will lose much efficiency in gaining



(a) (b) (c)

Figure 2: Regret Comparison between our methods and two benchmarks (RMLP-2 and Bandit). From the left to the right, the true underlying degree of smoothness is $m = \{2, 4, 6\}$ respectively. The x-axis denotes the time stamp T ranges from $1 \sim 12000$, and the y-axis denotes the regret at the time T defined in (6). We repeat the experiment 30 times and record the averaged regrets (solid lines) and standard errors (light areas) of every policy. The blue line denotes the regret of our policy (in Algorithm 1) with knowing degree of smoothness m and the orange line represents the regret of our policy (given in Algorithm 4) without knowing degree of smoothness m. The green and red lines are the regrets of implementing 'RMLP-2' and 'Bandit' policy respectvely.

revenue.

5.2.1 Real Application

Next, we leverage a simulation based on the real data to further illustrate the merits of our Algorithm over 'RMLP-2' and 'Bandit'.

We use the real-life auto loan dataset provided by the Center for Pricing and Revenue Management at Columbia University. This dataset is used by several related works [Phillips et al., 2015, Ban and Keskin, 2020, Luo et al., 2021, Wang et al., 2020a] and many others. The dataset contains $208,085$ auto loan applications received from July 2002 to November 2004. Some features such as the amount of loan, the borrower's information is contained in that dataset. We adopt the feature selection in the same way with Ban and Keskin [2020], Luo et al. [2021], Wang et al. [2020a] and consider the

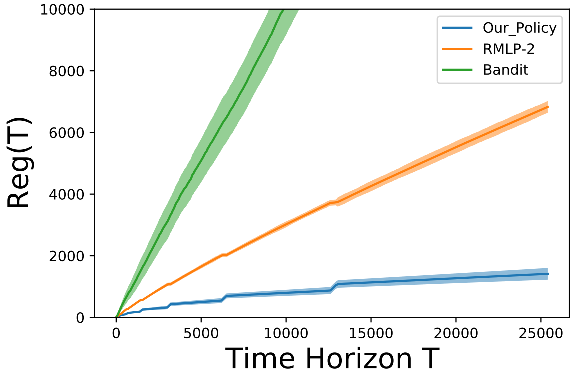


Figure 3: Comparison between our policy and 'RMLP-2' and 'Bandit' based on real data application.

following four features: the loan amount approved, FICO score, prime rate and competitor's rate. As for the price variable, we also computed it in the same way with the aforementioned literature, where $p_t = \text{Monthly Payment} \cdot \sum_{t=1}^{\text{Term}}(1 + \text{Rate})^{-t} - \text{Loan Amount}$. The rate is set as 0.12%, which is an approximate average of the monthly London interbank rate for the studied time period. Moreover, this dataset also records purchasing decision of the borrowers given the price set by the lender. For more details on this dataset, please refer to Phillips et al. [2015], Ban and Keskin [2020].

Note that one is not able to obtain online responses to any algorithms, thus, we follow the calibration idea proposed in Ban and Keskin [2020], Luo et al. [2021], Wang et al. [2020b] to first estimate the binary choice model and leverage it as the ground truth to conduct online numerical experiments. To be more specific, we first scale all variables into the scale of $[0, 1]$ (since the prediction results of single index model won't be affected by scale of the covariates). We randomly sample 5000 data points, estimate $\boldsymbol{\theta}_0$ and F using semi-parametric estimation tools from these data. We next treat them as the underlying true parameters for our binary choice model stated in (3). Given these key components, the remaining experiments remain almost the same as discussed in §5.1 and §5.2, except that here we set $\boldsymbol{\theta}_0$, distribution $F(\cdot)$ as the estimated one given above and sample $\mathbf{x}_t$ from those four features above. We set $B_0 = 4, \ell_0 = 200$ and conduct Algorithm 4 (in this algorithm, we use cross-validation to select m at the beginning of every episode, details are given in Algorithm 5).

We next compare Algorithm 4 with 'RMLP-2' and 'Bandit' policies. The details are given in Figure 3. To summarize, our policy outperforms the RMLP-2 [Javanmard and Nazerzadeh, 2019] and non-parametric bandit policy [Kleinberg and Leighton, 2003] in terms of both the regret performance and the ability to adapt to different noise distributions.

6 Conclusion

In this paper, we study the contextual dynamic pricing problem where the market value is linear in features, and the market noise has unknown distribution. We propose a policy that combines semi-parametric statistical estimation and online decision making. Our policy achieves near optimal regret, and is close to the regret lower bound where the market noise distribution belongs to a parametric class. We further generalize these results to the case when the product features satisfy the strong mixing condition. The practical performance of the algorithm is proved by extensive simulations.

There are several directions worth exploring in the future. First, we conjecture that the estimation accuracy of the market noise distribution F is crucial in the regret. Thus, within the function class $F \in \mathbb{C}^{(m)}$, we conjecture that a tighter regret lower bound $\Omega_d(T^{\frac{2m+1}{4m-1}})$ can be achieved instead of $\Omega_d(\sqrt{T})$, namely, our procedure is optimal. Second, in this work, we consider a linear model for the market value. In case a more complex model is appropriate, it's possible to extend our methodology to where the market value is nonlinear in product features, e.g. $v_t = \phi(\boldsymbol{\theta}_0^\top \mathbf{x}_t) + z_t$ or other structured statistical machine learning model such as the additive model $v_t = f_1(x_{t1}) + \cdots + f_d(x_{td}) + z_t$. Finally, it's worth studying similar pricing problems with adversarial or strategic buyers, which is potentially more suitable in some specific applications.

References

D. Alexey. Optimal non-parametric learning in repeated contextual auctions with strategic buyer. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR, 13–18 Jul 2020.

K. Amin, A. Rostamizadeh, and U. Syed. Repeated contextual auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Z. Anton and D. Alexey. Bisection-based pricing for repeated contextual auctions against strategic buyer. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11469–11480. PMLR, 2020.

M. Babaioff, S. Dughmi, R. Kleinberg, and A. Slivkins. Dynamic pricing with limited supply. *The ACM Transactions on Economics and Computation*, 2015. doi: 10.1145/2559152.

G. Ban and N. Keskin. Personalized dynamic pricing with machine learning: High dimensional features and heterogeneous elasticity. *Management Science*, 67(9):5549–5568, 2020.

A. Berg and D. Politis. Cdf and survival function estimation with infinite-order kernels. *Electronic Journal of Statistics*, 3:1436–1454, 2009.

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

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

R. J. Carroll, J. Fan, I. Gijbels, and M. P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997.

N. Cesa-Bianchi, T. Cesari, and V. Perchet. Dynamic pricing with finitely many unknown valuations. volume 98 of *Proceedings of Machine Learning Research*, pages 247–273. PMLR, 2019.

N. Chen and G. Gallego. Nonparametric pricing analytics with customer covariates. *Operations Research*, 69 (3):974–984, 2020.

Q. Chen, S. Jasin, and I. Duenyas. Nonparametric self-adjusting control for joint learning and optimization of multiproduct pricing with finite resource capacity. *Mathematics of Operations Research*, 44(2):601–631, 2019.

X. Chen, D. Simchi-Levi, and Y. Wang. Privacy-preserving dynamic personalized pricing with demand learning. *Management Science*, 2021.

X. Chen, Z. Owen, C. Pixton, and D. Simchi-Levi. A statistical learning approach to personalization in revenue management. *Management Science*, 68(3):1923–1937, 2022.

M. C. Cohen, I. Lobel, and R. Paes Leme. Feature-based dynamic pricing. *Management Science*, 66(11): 4921–5484, 2016.

M. Delecroix, W. Härdle, and M. Hristache. Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis*, 86(2):213–226, 2003.

A. V. den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*, 20(1):1–18, 2015.

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

A. V. den Boer and B. Zwart. Mean square convergence rates for maximum quasi-likelihood estimators. *Stochastic systems*, 4(2):375–403, 2015.

J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272, 1991.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. Chapman and Hall, 1996.

J. Fan and R. Li. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, 99(467):710–723, 2004.

J. Fan and Q. Yao. *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, 2003.

J. Fan, N. E. Heckman, and M. P. Wand. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 90(429):141–150, 1995.

N. Golrezaei, P. Jaillet, and J. C. N. Liang. Incentive-aware contextual pricing with non-parametric market noise. *arXiv:1911.03508*, 2019.

N. Golrezaei, A. Javanmard, and V. Mirrokni. Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Operations Research*, 69(1):297–314, 2020.

L. Györfi, A. Krzyżak, M. Kohler, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

P. G. Hall and J. S. Racine. Infinite order cross-validated local polynomial regression. *Journal of Econometrics*, 185(2):510–525, 2015.

W. Hardle, P. Hall, and H. Ichimura. Optimal Smoothing in Single-Index Models. *The Annals of Statistics*, 21 (1):157 – 178, 1993.

J. L. Horowitz. *Semiparametric methods in econometrics*, volume 131. Springer Science & Business Media, 2012.

J. L. Horowitz and W. Härdle. Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91(436):1632–1640, 1996.

H. Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1):71–120, 1993. ISSN 0304-4076.

A. Javanmard. Perishability of data: Dynamic pricing under varying-coefficient models. *The Journal of Machine Learning Research*, 18(53):1–31, 2017.

A. Javanmard and H. Nazerzadeh. Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research*, 20(1):315–363, 2019.

A. Javanmard, H. Nazerzadeh, and S. Shao. Multi-product dynamic pricing in high-dimensions with heterogeneous price sensitivity. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2652–2657, 2020.

N. B. Keskin and A. Zeevi. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research*, 26(5):1142–1167, 2014.

R. W. Klein and R. H. Spady. An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, 61:387–421, 1993.

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

P. R. Leme and J. Schneider. Contextual search via intrinsic volumes. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 268–282, 2018.

X. Li and Z. Zheng. Dynamic pricing with external information and inventory constraint. *Available at SSRN*, 2020.

A. Liu, R. P. Leme, and J. Schneider. Optimal contextual pricing and extensions. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1059–1078, 2021.

Y. Luo, W. W. Sun, and Y. Liu. Distribution-free contextual dynamic pricing. *arXiv:2109.07340*, 2021.

Y. P. Mack and B. W. Silverman. Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 61:405–415, 1982.

B. K. Mallick and A. E. Gelfand. Generalized linear models with unknown link functions. *Biometrika*, 81(2): 237–245, 1994.

J. Mao, R. Leme, and J. Schneider. Contextual pricing for lipschitz buyers. In *Advances in Neural Information Processing Systems*, volume 31, pages 5643–5651, 2018.

E. Masry. Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, 17:571–599, 1996.

C. E. McCulloch. Generalized linear models. *Journal of the American Statistical Association*, 95(452):1320–1324, 2000.

L. T. McMurry and N. D. Politis. Nonparametric regression with infinite order flat-top kernels. *Journal of Nonparametric Statistics*, 16(3-4):549–562, 2004.

S. Miao, X. Chen, X. Chao, J. Liu, and Y. Zhang. Context-based dynamic pricing with online clustering. *ArXiv:1902.06199*, 2019.

V. V. Misic and G. Perakis. Data analytics in operations management: A review. *Manufacturing & Service Operations Management*, 22(1):158–169, 2020.

E. A. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9(1):141–142, 1964.

M. Nambiar, D. Simchi-Levi, and H. Wang. Dynamic learning and pricing with model misspecification. *Management Science*, 65(11):4980–5000, 2019.

R. L. Phillips, A. S. Simsek, and G. J. v. Ryzin. The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Science*, 61:1741–1759, 2015.

J. L. Powell, J. H. Stock, and T. M. Stoker. Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, 57(6):1403–1430, 1989.

S. Qiang and M. Bayati. Dynamic pricing with demand covariates. *Stochastic Models eJournal*, 2016.

D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric regression*. Cambridge university press, 2003.

P. Rusmevichientong, B. Van Roy, and P. W. Glynn. A nonparametric approach to multiproduct pricing. *Operations Research*, pages 82–98, 2006.

V. Shah, R. Johari, and J. Blanchet. Semi-parametric dynamic contextual pricing. In *Advances in Neural Information Processing Systems*, volume 32, pages 2363–2373, 2019.

B. W. Silverman. Weak and Strong Uniform Consistency of the Kernel Estimate of a Density and its Derivatives. *The Annals of Statistics*, 6(1):177 – 184, 1978.

C. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980.

C. J. Stone. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10 (4):1040 – 1053, 1982.

W. Tang, C.-J. Ho, and Y. Liu. *Differentially Private Contextual Dynamic Pricing*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2020.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008. ISBN 0387790519.

C.-H. Wang, Z. Wang, W. W. Sun, and G. Cheng. Online regularization for high-dimensional dynamic pricing algorithms. *arXiv preprint arXiv:2007.02470*, 2020a.

Y. Wang, X. Chen, X. Chang, and D. Ge. Uncertainty quantification for demand prediction in contextual dynamic pricing. *Production and Operations Management*, 30, 12 2020b.

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

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A.*, 26(4):359–372, 1964.

M. M. Wei and F. Zhang. Recent research developments of strategic consumer behavior in operations management. *Computers and Operations Research*, 93:166–176, 2018. ISSN 0305-0548.

S. Weisberg and A. H. Welsh. Adapting for the missing link. *The Annals of Statistics*, 22(4):1674–1700, 1994.

Y. Xia and W. K. Li. On single-index coefficient regression models. *Journal of the American Statistical Association*, 94(448):1275–1285, 1999.