Policy Optimization Using Semiparametric Models for Dynamic Pricing

Jianqing Fan, Yongyi Guo, and Mengxin Yu

Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ

ABSTRACT
In this article, 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 the work by Javanmard and Nazerzadeh 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 cdf $F(\cdot)$ with $m$th order derivative ($m \geq 2$), our policy achieves a regret upper bound of $\tilde{O}(d(T^{m+1}))$, where $T$ is the time horizon and $\tilde{O}(\cdot)$ is the order hiding logarithmic terms and the feature dimension $d$. The upper bound is further reduced to $\tilde{O}(d(\sqrt{T}))$ if $F$ is super smooth. These upper bounds are close to $\Omega(d(T^{4/3}))$, 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. Supplementary materials for this article are available online.

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 $t = 1, \ldots, T$. Each product is attached with a known feature vector $x_t \in \mathbb{R}^d$. In addition, the product’s market value $v_t$ is linear in the features plus some iid market noise $z_t$ with an unknown cumulative distribution $F(\cdot)$:

$$v_t = \theta_0^\top x_t + z_t, \quad z_t \sim F.$$  

Here $\tilde{x}_t = (x_t^\top, 1)^\top$ and $\theta_0$ 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 $P(v_t \geq p_t)$ actually depends on both the parameter $\theta_0$ as well as the distribution of $z_t$, which admits a semiparametric 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 uses semi-parametric estimation techniques to achieve a low regret. In particular, under mild regularity conditions, if the cdf $F(\cdot)$ of $z_t$ has $m$th derivative, the regret over a time horizon $T$ is upper bounded by $\tilde{O}((Td)^{m+1} \log(T + \log(T/d)))$, where $d$ is the number of features. This result is further generalized to a setting where the product features $x_t$ are not independent, as long as $\{x_t\}_{t \geq 1}$ is a stationary series that satisfies certain $\beta$-mixing conditions. Moreover, when $F$ is infinitely differentiable, the total regret can be upper bounded by $\tilde{O}((Td)^{2m+1} / (\log T)^{m+3} \log(d + 1) + \log(T/d))$. This rate is the same as the parametric lower bound up to some logarithmic factors, that is, 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 nonparametric 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, Van Roy, and Glynn (2006), Besbes and Zeevi (2009), Broder and Rusmevichientong (2012), Keskin and Zeevi (2014), den Boer and Zwart (2014), Wang, Deng, and Ye (2014), den Boer and Zwart (2015), Babaioff et al. (2015), Cesa-Bianchi, Cesari, and Perchet (2019), and Chen, Jin, and Duenyas (2019). As an example, Cesa-Bianchi, Cesari, and Perchet (2019) study the dynamic pricing problem where the buyer’s valuation of a product is supported on a finite set of 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 $O(K \log T)$. 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 value of the product at time $t$ as a linear function of its features $x_t$ plus some market noise $z_t$, that is, $v_t = \theta^T x_t + z_t$ where $\theta$ 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, Simchi-Levi, and Wang 2021; Tang, Ho, and Liu 2020; Golrezaei, Javanmard, and Mirrokni 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 $x_t$, 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 $O_d(\log T)$, where $O_d$ 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 $O_d(\log T)$ to $O_d(\sqrt{T})$. Other related works include Chen, Simchi-Levi, and Wang (2021) and Tang, Ho, and Liu (2020) where the authors explore certain differentially private policies under similar model setting; Golrezaei, Javanmard, and Mirrokni (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 $\theta$ 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 Section 2, this leads to modeling the demand curve as a generalized linear model with unknown link, and will be our main focus in this article. 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 $O_d(T)$. Golrezaei, Javal, and Liang (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 $O_d(\sqrt{T})$. Golrezaei, Javanmard, and Mirrokni (2020) also explore the second price auction and derive a regret upper bound of $O_d(T^{2/3})$ compared to a “robust benchmark” where the price maximizes the revenue of the worst link function in the class. Shah, Johari, and Blanchet (2019) explore an alternative setting where the market value $v_t = \exp(\theta^T x_t + z_t)$ and $z_t$ has unknown distribution. By using this specific structure, the authors propose the DEEP-C algorithm based on multi-arm bandit that has a regret upper bound of $O_d(\sqrt{T})$. The authors also propose some variants of the algorithm and study them via simulations. Recently, Luo, Sun, and Liu (2021) study a similar problem to ours, assuming a linear market valuation with unknown noise distribution. They provide a DIP policy that achieves regret $O_d(T^{2/3} + \|\theta - \theta^*\|_1 T)$, where $\|\theta - \theta^*\|_1$ is the estimation accuracy of the parameter $\theta$.

There are some literature studying other dynamic pricing algorithms (Amin, Rostamizadeh, and Syed 2014; Cohen, Lobel, and Paes Leme 2016; Mao, Leme, and Schneider 2018; Leme and Schneider 2018; Nambiar, Simchi-Levi, and Wang 2019; Anton and Alexey 2020; Alexey 2020; Ban and Keskin 2020; Li and Zheng 2020; Javanmard, Nazerzadeh, and Shao 2020; Chen and Gallego 2020; Liu, Leme, and Schneider 2021). For example, Mao, Leme, and Schneider (2018) study a nonparametric dynamic pricing pricing where the market value is modeled as a general nonparametric function $f(x_t)$, where $x_t$ are the features. A binary feedback is similarly observed based on the comparison between $f(x_t)$ and the proposed price. The authors apply a variation of midpoint algorithm and achieve a regret upper bound of $O(T^{d/(d+1)})$ with $d$ being the dimension of $x_t$.

Semi-parametric and nonparametric 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, Stock, and Stoker 1989; Ichimura 1993; Hardle, Hall, and Ichimura 1993; Klein and Spady 1993; Weisberg and Welsh 1994; Mallick and Gelfand 1994; Horowitz and Hardle 1996; Carroll et al. 1997; Xia and Li 1999; Delecroix, Hardle, and Hristache 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, Wand, and Carroll (2003), Tsybakov (2008), and Horowitz (2012). Various methods have been proposed to estimate the parametric part that achieves root-$n$ consistency under certain conditions (Powell, Stock, and Stoker 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 nonparametric 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 nonparametric regression techniques can be used to estimate the nonparametric part of the single index model that achieves $\ell_\infty$ 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 article, we'll provide nonasymptotic error bounds for both coefficient estimation as well as the plug-in Nadaraya-Watson estimator in a uniform sense. These nonasymptotic 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 $F \in C^2$ where $F$ is the cdf of $z_t$, the regret over a time horizon $T$ is upper bounded by $\tilde{O}(T^{d+1}/m)$; if $F$ is “super smooth,” the bound is further reduced to $\tilde{O}(\sqrt{Td})$, 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 $\beta_0$ is imposed. Table 1 illustrates the settings of our work as well as several related literatures. Golrezaei, Javanmard, and Mirrokni (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, Johari, and Blanchet (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 $\tilde{O}(\sqrt{Td}^{1/4})$. 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 $O(\sqrt{T})$ regret. Our work partly answers their guess by providing a policy with a $\tilde{O}(\sqrt{Td})$ regret when the demand function is sufficiently smooth. Compared to the DIP policy in Luo, Sun, and Liu (2021) and its regret $O(T^{2/3} + ||\hat{\theta} - \theta||_1 T)$, we are more clear on how $\theta$ is estimated within the pricing algorithm, and we provide explicit rate on both the estimation error and the regret. Moreover, compared to several fully nonparametric dynamic pricing literatures, such as Mao, Leme, and Schneider (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 $\{x_t\}_{t \geq 1}$ 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 $\ell_\infty$ 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 nonparametric 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, supplementary materials for related lemmas.

### 1.3. Notation

Throughout this work, we use $[n]$ to denote $\{1, 2, \ldots, n\}$. For any vector $x \in \mathbb{R}^n$ and $q \geq 0$, we use $\|x\|_q$ to represent the vector $\ell_q$ norm, that is, $\|x\|_q = (\sum_{i=1}^n |x_i|^q)^{1/q}$. In addition, we let $\nabla_x L(\cdot), \nabla^2_x L(\cdot)$ be the gradient vector and Hessian matrix of
loss function $L(\cdot)$ with respect to $x$. For any given matrix $X \in \mathbb{R}^{d_1 \times d_2}$, we use $\| \cdot \|$ to denote the spectral norm of $X$ and we write $X \succ 0$ or $X \preceq 0$ if $X$ or $-X$ is semidefinite. For any event $A$, we let $\mathbb{1}_A$ 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 $(a_n)_{n \geq 1}$, $(b_n)_{n \geq 1}$, we write $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ if there exists a positive constant $C$ such that $a_n \leq C \cdot b_n$ and we write $a_n = o(b_n)$ or $a_n \ll b_n$ if $a_n/b_n \to 0$. In addition, we write $a_n = \Omega(b_n)$ or $a_n \gtrsim b_n$ if $a_n/b_n \geq c$ 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 $\mathcal{O}(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ to denote similar meanings as above while treating the variable $d$ as fixed. Moreover, we let $\mathcal{O}(\cdot)$, $\mathcal{O}(\cdot)$, $\mathcal{O}(\cdot)$ 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 article is organized as follows. We describe the problem in Section 2 and propose a solution in Section 3 where some heuristic arguments are offered for bounding the regret. In Section 4, we provide our theoretical results on the upper bounds of the regret and in Appendix C, supplementary materials, we discuss a lower bound result. Our algorithm is illustrated in Section 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 $t = 1, 2, \ldots, T$. 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 $v_t \in [\beta_t, B - \delta_t]$ for some constant $\delta_t > 0$. At each period $t$, the seller posts a price $p_t$. If $p_t \leq v_t$, 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

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

(1)

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

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

$$v_t = \theta_0^\top \tilde{x}_t + z_t,$$

(2)

where $\theta_0 = (\beta_0^\top, \alpha_0^\top) \in \mathbb{R}^{d+1}$ is an unknown parameter, and $\{z_t\}_{t \geq 1}$ is an iid 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

$$y_t = \begin{cases} +1 & \text{with probability } 1 - F(p_t - \theta_0^\top \tilde{x}_t), \\ 0 & \text{with probability } F(p_t - \theta_0^\top \tilde{x}_t). \end{cases}$$

(3)

Remark 1. In fact, each $x_t$ 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 $v_t \geq 0$, which is more reasonable in practice (Otherwise $v_t$ has positive probability to be negative, since $z_t$ is independent with the covariates $x_t$).

In a nondynamic 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, Heckman, and Wand 1995; Carroll et al. 1997). In their works, it’s usually assumed that $p_t = 0$ and $(\{x_t\})_{t \geq 1}$ are independent observations, and the goal is to estimate $\theta_0$ and $F(\cdot)$.

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 $\theta_0$ and $F(\cdot)$. These two problems are closely related but also decisively different.

We now state our objective in more details. Given observed features $x_t$, the expected revenue at time $t$ with a posted price $p$ is

$$\text{rev}_t(p, \theta_0, F) := \mathbb{E} p \cdot \mathbb{I}(v_t \geq p) = p(1 - F(p - \theta_0^\top \tilde{x}_t)).$$

(4)

The optimal posted price $p^*_t$ for a product with attribute $x_t$ is given by

$$p^*_t = \arg\max_{p \geq 0} \{p(1 - F(p - \theta_0^\top \tilde{x}_t))\},$$

(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 $\pi$ that governs the rule of posted prices $\{p_t\}_{t \geq 1}$ by investigating the regret compared to the “oracle pricing policy” that uses the knowledge of both $\theta_0$ 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

$$\text{Regret}_{\pi}(T) \equiv \max_{\theta_0 \in \mathcal{Q}(x), \pi(\cdot)} \mathbb{E} \left[ \sum_{t=1}^{T} \left( p^*_t \mathbb{I}(v_t \geq p^*_t) - p_t \mathbb{I}(v_t \geq p_t(\pi)) \right) \right],$$

(6)

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

Apparently, learning $\theta_0$ 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 $c_\phi$ such that $\phi'(u) \geq c_\phi$ for all $u \in (-\delta, \delta)$, 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

$$p_t^* = g(\theta_0^T \mathbf{x}_t),$$

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

**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 $\mathbf{x}_t$ and unknown but fixed $\theta_0$. 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 $c_\phi \geq 1$.

---

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

1. **Input:** Upper bound of market value (\{v_t\}_{t=1}^\infty): $B > 0$, minimum episode length: $\ell_0$, degree of smoothness: $m$.

2. **Initialization:**
   
   $p_1 = 0$, $\hat{\theta}_1 = 0$.

3. **for** each episode $k = 1, 2, \ldots$ **do**

4. **Set** length of the $k$th episode $\ell_k = 2^{k-1} \ell_0$; Length of the exploration phase $a_k = [\ell_k, d_{m+1}]$.

5. **Exploration Phase** ($t \in I_k := [\ell_k, \ldots, \ell_k + a_k - 1]$):

   - **Offer price** $p_t \sim \text{Unif}(0, B)$.

6. **Updating Estimates at the end of the exploration phase** with data [(\mathbf{x}_t, y_t)]_{t \in I_k}:

   - **Update** estimate of $\theta_0$ by $\hat{\theta}_k = \hat{\theta}_k((\mathbf{x}_t, y_t))_{t \in I_k}$:

     $$\hat{\theta}_k = \arg\min_\theta L_k(\theta) := \frac{1}{|I_k|} \sum_{t \in I_k} (y_t - \theta^T \mathbf{x}_t)^2 \quad (7)$$

7. **Update** estimates of $F$, $F'$ by $F_k(u, \hat{\theta}_k) = F_k(u; \hat{\theta}_k, ([\mathbf{x}_t, y_t, p_t])_{t \in I_k})$, $F_k(u, \hat{\theta}_k) = F_k(u; \hat{\theta}_k, ([\mathbf{x}_t, y_t, p_t])_{t \in I_k})$.

   - The detailed formulas are given by (14) and (16).

8. **Update** estimate of $\phi$ by $\hat{\phi}_k(u) = u - \frac{1-F_k(u)}{F_k(u)}$ and estimate of $g$ by $\hat{g}_k(u) = u + \hat{\phi}_k^{-1}(-u)$.

9. **Exploitation Phase** ($t \in I_k := [\ell_k + a_k, \ldots, \ell_k + 1]$):

   - **Offer** $p_t$ as

   $$p_t = \min\{\max[\hat{g}_k(\mathbf{x}_t^T \hat{\theta}_k), 0], B\}. \quad (8)$$

---

### 3. Algorithm and Basic Regret Analysis

We first propose Algorithm 1 in Section 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 Section 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 iid to collect the data and model parameters (i.e., $\theta$, $F$) 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 $\mathbf{x}_t$. The details are stated in Algorithm 1.

Despite semiparametric model (3) with unknown link, by offering $p_t \sim \text{Unif}(0, B)$, $\mathbf{y}_t$ follows the linear model with regression $\mathbf{x}_t^T \theta_0$ and this leads to the least-squares estimate (7). To see this, it follows that

$$\mathbb{E}[\mathbf{y}_t | \mathbf{x}_t] = \mathbb{E}\mathbf{y}_t|\theta_0 = \mathbb{E}v_t$$

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

$$\mathbb{E}v_t = \mathbb{E}\mathbb{E}[v_t | \mathbf{x}_t, z_t] = \mathbb{E}\mathbb{E}[v_t(1 - p_t) | \mathbf{x}_t, z_t] = \mathbb{E}F_p(\mathbf{x}_t^T \theta_0 + z_t) f_1(1) = f_1(1) + f_1(1 - F_p(\mathbf{x}_t^T \theta_0 + z_t)) f_1(0)$$

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 $B_k = C/\log|I_k|$ such that $v_t$ falls in $(-B_k, B_k)$ with high probability. We then offer $p_t \sim \text{Unif}(-B_k, B_k)$. Conditional on $v_t \in (-B_k, B_k)$, $B_k(2B_k - 1)$ serves as an unbiased estimator for $\mathbf{x}_t^T \theta_0$. 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) \( \ell \) is known and \( d \) is bounded.

First, denote \( \ell_1 \) 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

\[
R_1 = O(\ell_1).
\]

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

\[
R_t := p_t^*(\hat{x}(t \geq p_t^*)) - p_t(\hat{x}(t \geq p_t)).
\]

Then the conditional expectation of regret at time \( t \) given previous information and \( \hat{x}_t \) is

\[
E[R_t | \hat{x}_{t-1}] = E[p_t^*(\hat{x}(t \geq p_t^*)) - p_t(\hat{x}(t \geq p_t)) | \hat{x}_{t-1}]
\]

\[
= p_t^*(1 - F(p_t^* - \hat{x}(t \theta_0))) - p_t(1 - F(p_t - \hat{x}(t \theta_0))
\]

\[
= \text{rev}(p_t^*, \theta_0, F) - \text{rev}(p_t, \theta_0, F)
\]

\[
(10)
\]

Here \( \hat{x}_t = \sigma(x_1, x_2, \ldots, x_{t+1}; z_1, \ldots, z_t) \). 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

\[
E[R_t | \hat{x}_{t-1}] \leq (p_t - p_t^*)^2 = (\hat{g}(\hat{\theta} \hat{x}_t) - g(\hat{\theta} \hat{x}_t))^2
\]

\[
\leq 2(\hat{g}(\hat{\theta} \hat{x}_t) - g(\hat{\theta} \hat{x}_t))^2 + 2(\hat{g}(\hat{\theta} \hat{x}_t) - g(\theta_0 \hat{x}_t))^2
\]

\[
\approx J_t + J_2.
\]

In fact, \( J_2 \) is upper bounded by \( \|\hat{\theta} - \theta_0\|_2^2 \) (given the Lipschitz property of \( g \) according to Assumption 1 and suitable conditions over \( P_X \)). By solving (7), we prove that the squared \( \ell_2 \) error is of order \( O(\ell_1^{-1}) \), which is the order of \( J_2 \). The term \( J_1 \) is upper bounded by \( \|\hat{g} - g\|_{\infty} \), and is further bounded by \( \max\{\|\hat{F} - F\|_{\infty}, \|\hat{F} - F\|_{\infty}^2\} \). Note that by (1), \( F(\cdot) \) is the nonparametric function of 1 \(-\) Y1 given \( w_t = p_t - \hat{x}_t \theta_0 \), in which \( p_t \) is the observed price given in the exploration phase. Since \( \theta_0 \) is estimated at a faster rate, we can assume that \( w_t \) is observable given a proper estimator of \( \theta_0 \). Therefore, the error rate is dominated by estimating \( F(\cdot) \). Assuming \( F \) has an \( m \)th continuous derivative, we construct \( \hat{g} \) using the kernel estimator with an \( m \)th order kernel, and prove that \( \max\{\|\hat{F} - F\|_{\infty}, \|\hat{F} - F\|_{\infty}^2\} \leq O(\ell_1^{-(m-1)/(2m-1)}) \) in which a logarithmic order is ignored for simplicity of presentation. Therefore, the total regret during the exploration phase can be upper bounded by

\[
R_2 \leq \ell \cdot \ell_1^{-2(m-1)/(2m-1)}.
\]

Combining (9) and (12), we know that by choosing \( \ell_1 \) of the order of \( \ell^{2(m-1)/(4m-1)} \), we balance the regret of both exploration and exploitation phases, and the total regret during the episode is given by

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

For a second order kernel, the above regret is of order \( O(\ell^{5/7}) \). For a relatively large \( m \), the regret is close to \( O(\ell^{1/2}) \), 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 Section 4.1, we consider the setting with independent covariates and finite differentiable noise distributions. In Section 4.2, we further extend our results in Section 4.1 to the setting with correlated features. Finally, we extend the aforementioned results to the regime with infinitely differentiable noise distributions that is, \( m = \infty \) in Section 4.3.

#### 4.1. Result under Independence Settings

The main result of this section is Theorem 1. To obtain this result, 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 interesting of their own rights.

Assume that \( \|\theta_0\| \leq R_0 \) for some constant \( R_0 > 0 \). We also define \( R_X := \sup_{x \in X} \|x\|_2 \). Before stating our main results, we first make the following assumptions on \( x_t \).

**Assumption 2.** There exist positive constants \( c_{\min} \) and \( c_{\max} \), such that the covariance matrix \( \Sigma \) given by \( \Sigma = E[\hat{x}_t \hat{x}_t^T] \) satisfies \( c_{\min} \| \Sigma \| \leq \| \Sigma \| \leq c_{\max} \| \Sigma \| \), where \( \hat{x}_t = (x_t^T, 1)^T \).

As we observe from (11), bounding the regret in the exploitation phase needs to estimate both parameter \( \theta_0 \) and function \( g(\cdot) \). In the following, we first present an upper bound of estimating \( \theta_0 \) 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 \( |I_k| \geq c_0 (d + 1) \), with probability at least \( 1 - 2e^{-c_1 c_{\min} |I_k|/16} - 2/|I_k| \),

\[
|\hat{\theta}_k - \theta_0\|_2 \leq \frac{8 \max(R_X, 1)(R_X R_0 + B)}{c_{\min} \sqrt{(d + 1) \log |I_k| / |I_k|}}.
\]

(13)

Let \( \Theta_k := \mathcal{B}(\theta_0, R_k) \), 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 \( J_2 \) for any episode \( k \).

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

We use the kernel regression method and \( \hat{\theta}_k \) obtained above to construct \( \hat{F}_k \) and \( \hat{F}'_k \). Recall that by (3), we have \( E(y_t | w_t(\theta_0)) = 1 - F(w_t(\theta_0)) \) where \( w_t(\theta) := p_t - \hat{x}_t^T \theta \). Recall \( p_t \) is the observed price offered in the \( k \)th exploration phase. Thus, given \( \hat{\theta}_k \), \( F(\cdot) \) 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
\[
\hat{F}_k(u, \theta) = 1 - \hat{\gamma}_k(u, \theta) = 1 - \frac{h_k(u, \theta)}{f_k(u, \theta)},
\]
and \(\hat{F}_k(u) = \hat{F}_k(u, \hat{\theta}_k)\), where
\[
h_k(u, \theta) = \frac{1}{|I_k|b_k} \sum_{i \in I_k} K\left(\frac{w_i(\theta) - u}{b_k}\right) Y_i, \\
f_k(u, \theta) = \frac{1}{|I_k|b_k} \sum_{i \in I_k} K\left(\frac{w_i(\theta) - u}{b_k}\right),
\]
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, \(\hat{F}_k^{(1)}(u) = \hat{F}_k^{(1)}(u, \hat{\theta}_k)\) where
\[
\hat{F}_k^{(1)}(u, \theta) = -\hat{\gamma}_k^{(1)}(u, \theta) \\
= -\frac{h_k^{(1)}(u, \theta)f_k(u, \theta) - h_k(u, \theta)f_k^{(1)}(u, \theta)}{f_k^2(u, \theta)},
\]
\[
h_k^{(1)}(u, \theta) = -\frac{1}{|I_k|b_k^2} \sum_{i \in I_k} K'\left(\frac{w_i(\theta) - u}{b_k}\right) Y_i, \\
f_k^{(1)}(u, \theta) = -\frac{1}{|I_k|b_k^2} \sum_{i \in I_k} K'\left(\frac{w_i(\theta) - u}{b_k}\right).
\]

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

Assumption 4. \(r_{\theta}(u) := \mathbb{E}[y_i | w_i(\theta) = u]\) satisfies the following:

- (Smoothness) \(h_{\theta}(u) = f_{\theta}(u)r_{\theta}(u) \in C^m, h_{\theta}^{(m)}\) is \(l_{\gamma}\)-Lipschitz on \(I\) for all \(\theta \in \Theta_0\). Here \(m\) and \(l_{\gamma}\) are defined in Assumption 3.
- (Lipschitz) There exists a constant \(l_{\theta}\) such that \(r_{\theta_0} = 1 - F\) is \(l_{\gamma}\)-Lipschitz, and for any \(\epsilon > 0\), \(\sup_{|\theta - \theta_0| \leq \epsilon, u \in I} |r_{\theta}(u) - r_{\theta_0}(u)| \leq l_{\gamma}\epsilon\).

Assumption 5. The kernel \(K\) satisfies the following:

- (Order-\(m\) kernel) \(\int K(s)ds = 1, \int s^j K(s)ds = 0\) for \(j \in \{1, \ldots, m - 1\}\), and \(\int |s|^m K(s)ds < +\infty\). Here \(m\) is the same as in Assumption 3.
- (Lipschitz) Both \(K(s)\) and \(K'(s)\) are \(l_{\gamma}\)-Lipschitz continuous with bounded support.

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

Remark 7. Assumptions 3 and 4 can be relaxed in terms of the smoothness requirements: For all \(m \geq 3\), we only need \(f_{\theta}(u), h_{\theta}(u) \in C^{m-1}\), and that \(f_{\theta}^{(m-1)}(u), h_{\theta}^{(m-1)}(u)\) are \(\ell\)-Lipschitz for some constant \(\ell\). For \(m = 2\), we only need \(f_{\theta}(u), h_{\theta}(u) \in C^{1}\), and that the second order derivatives of \(f_{\theta}(u), h_{\theta}(u)\) exist and are bounded. One is able to see assuming functions in \(C^m\) 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 \(\ell\)-Lipschitz continuous (i.e., may it not be differentiable), we also provide an alternative algorithm in Appendix H, supplementary materials which achieves a regret upper bound \(O(T^{3/4})\).

Remark 9. One is also able to estimate \(F(u)\) and \(F'(u)\) 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 \([-\delta_2, \delta_2]\) satisfy Assumption 4.2. The proof is very similar. For simplicity, we only focus on studying kernel regression in this article.
Lemma 3. Under the same conditions as Lemma 2, with probability at least $1 - 4\delta$, we have

$$
\sup_{u \in [\delta_1, B - \delta_2]} |\widehat{g_k}(u) - g(u)| \\
\leq \bar{C}_{x,K} |I_K|^{-\frac{m-1}{m+1}} \sqrt{\log |I_K|/(\sqrt{d} + \sqrt{\log \frac{1}{\delta}})}.
$$

(19)

We next develop a uniform upper bound for term $J_1$ given in (11) for the $k$th episode in Lemma 4.

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

$$
\sup_{u \in [\delta_1, B - \delta_2]} |\widehat{g_k}(u) - g(u)| \\
\leq \bar{C}_{x,K} |I_K|^{-\frac{m-1}{m+1}} \sqrt{\log |I_K|/(\sqrt{d} + \sqrt{\log \frac{1}{\delta}})}.
$$

Remark 10. In Algorithm 1 we define $\widehat{g_k}(u) = u + \phi_k^{-1}(-u)$ with $u \in [\delta_1, B - \delta_2]$. Thus, computing $\widehat{g_k}(u)$ involves obtaining the inverse of $\phi_k$, which is not necessarily monotonic. Nevertheless, it’s not difficult to define or compute $\phi_k^{-1}$. In fact, we’ll show in the proof of Lemma 4 that $\phi_k$ is very “close” to $\phi$ in some main interval of interest, which contains $[\phi^{-1}(\delta_2 - B), \phi^{-1}(-\delta_2)]$ and depends only on $F$. (Recall in Assumption 1 that $\phi'$ is bounded below from 0, so $\phi$ is strictly increasing. Thus, for any $u \in [\delta_1, B - \delta_2]$, the above fact will guarantee the existence of $\phi_k^{-1}(-u)$ as some $x$ within the interval such that $\phi_k(x) = -u$.

Combining the above lemmas, which give us upper bounds for terms $J_1, J_2$ 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 $B_{x,K}$, $B_{x,K}^*$, and $C_{x,K}$ (depending only on the absolute constants within the assumptions) such that for all $T$ satisfying

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

the regret of Algorithm 1 over time $T$ is no more than $C_{x,K}^*(\tilde{O}(T)^{\frac{m-1}{m+1}} \log T(1 + \log T/d))$.

Remark 11. We note that Golrezaei, Javanmard, and Mirroroki (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 $p^*_k$ from (5) with true $\theta_0$ and $F()$. In Golrezaei, Javanmard, and Mirroroki (2020), the authors instead set the benchmark $p^*_k$ so as to maximize the worst function in their function class $F$, that is,

$$
p^*_k = \arg\max_{p \succeq 0} \min_{\theta \in \Theta} \rho(1 - F(p - \theta^\top \tilde{x}_i)).
$$

Their optimal regret is of order $O_d(T^{2/3})$, while ours is $O_d(T^{m+1})$, which is closer to $O_d(T^{1/2})$ 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 semiparametric setting.

On the other hand, Golrezaei, Jaillet, and Liang (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 $F()$. A popular choice in non-parametric 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 $k \geq 2$, 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 Appendix I, supplementary materials.

- **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 performances of the algorithm ($\tilde{O}((Td)^{5/7})$ and $O((Td)^{3/5})$) 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 article 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 $x_t$ can be dependent. For this purpose, we first impose the strong-mixing condition which measure the dependence between covariates over time.

**Definition 1.** [$\beta$-mixing] For a sequence of random vectors $x_t \in \mathbb{R}^{d \times 1}$ on a probability space $(\Omega, \mathcal{A}', \mathbb{P})$, define $\beta$-mixing coefficient

$$
\beta_k = \sup_{l \geq 0} \beta(\sigma(x_t, t \leq l), \sigma(x_t, t \geq l + k))
$$

in which

$$
\beta(A, B) = \frac{1}{2} \sup_{\{i \in I\}} \left\{ \sum_{j \in J} \mathbb{E}\left[ (\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j))^2 \right] \right\},
$$

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

The following assumption ensures that $\{x_t\}_{t \geq 1}$ are not too strongly dependent. Combining with other assumptions, we ensure that the empirical covariance matrix $\frac{1}{T} \sum_{t=1}^{T} x_t x_t^\top$ concentrate around the population version, which is necessary in deriving the regret in every episode.
Assumption 6. The sequence $x_t, t \geq 0$ are strictly stationary time series and follow a joint density of $\theta$ within a certain class. We assume that $k \leq e^{-cT}$ holds with some constant $c$.

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

Assumption 7. Let $\rho_i(u_i, u_j) := \mathbb{E}[w_i(u_i) | \theta] = u_i$, $i \geq 0$, $\theta_0(u_i) := \mathbb{E}[w_i(u_i) | \theta] = u_i$, $i \geq 0$ be the joint regression function and marginal regression function. In addition, we also set $f_0(u_i, u_j)$, $i \geq 0$, $\theta_0(u_i)$, $i \geq 0$ as the joint density of $w_i(\theta)$ and $w_i(\theta)$ and marginal density of $w_0(\theta)$, respectively. Then we define $g_{ij}(u_i, u_j) := \rho_i(u_i, u_j)f_i(u_i, u_j) - \rho_i(u_i)f_i(u_j)\theta_0(u_i)f_i(u_j)$ and $g_{ij}(u_i, u_j) = \rho_i(u_i, u_j) - \rho_i(u_i)f_i(u_j)$. We assume $g_{ij}(u_i, u_j)$ and $g_{ij}(u_i, u_j)$ follow $i$-Lipschitz continuous condition, in a sense that

$$|g_{ij}(u_i, u_j) - g_{ij}(u_i', u_j')| \leq \sqrt{(u_i - u_i')^2 + (u_j - u_j')^2}, \quad q \in [1, 2]$$

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

When the covariates $x_t, x_s$ are independent, we can get $g_{ij}(u_i, u_j) = 0, q \in [1, 2]$, for all $(u_i, u_j)$. Under such a mild assumption, we obtain a uniform upper bound of $|g_{ij}(u_i, u_j)|$, which is dominated by the $\beta$-mixing constant $\beta_{ij}, \forall i, j$.

Remark 13. This assumption is quite standard, and ensures that $f_0(u), f_0(u) \in C^\infty$. The class of functions is still infinite dimensional. However, the class of super-smooth functions has been used in nonparametric density literature. In particular, it has been used in Fan (1991) for characterizing the difficulty of nonparametric deconvolution.

Assumption 8. Define $\phi_\theta, \xi_\theta, \phi_\theta^{(1)}, \xi_\theta^{(1)}$ as the Fourier transform of the function $f_0, h_0, f_0^{(1)}$, and $h_0^{(1)}$, respectively:

$$\phi_\theta(s) = \int_{-\infty}^{\infty} f_0(x)e^{isx}dx, \quad \xi_\theta(s) = \int_{-\infty}^{\infty} h_0(x)e^{isx}dx,$$

$$\phi_\theta^{(1)}(s) = \int_{-\infty}^{\infty} f_0^{(1)}(x)e^{isx}dx, \quad \xi_\theta^{(1)}(s) = \int_{-\infty}^{\infty} h_0^{(1)}(x)e^{isx}dx,$$

and $h(x) = f_0(x)\theta_0(x)$. There exist positive constant $D_{\theta}$ and $d_{\theta}$ and $\alpha > 0$ such that

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

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

Lemma 5. Under Assumption 8, there exists a positive constant $C_{\inf}$ depending only on $\alpha, D_{\theta}$, and $d_{\theta}$ such that for all kernel $K$ satisfying (20), for each episode $k$, by choosing the bandwidth $b_k = c_\epsilon(d_{\theta}/\log |I_k|^{1/2}$ in (15) and (17), we have

$$\sup_{\theta \in \Theta_0} |E[f_0(u, \theta)] - f_0(u)| \leq C_{\inf} (d_{\theta}/\log |I_k|^{1/2}),$$

$$\sup_{\theta \in \Theta_0} |E[h_0(u, \theta)] - h_0(u)| \leq C_{\inf} (d_{\theta}/\log |I_k|^{1/2}),$$

$$\sup_{\theta \in \Theta_0} |E[f_0^{(1)}(u, \theta)] - f_0^{(1)}(u)| \leq C_{\inf} (d_{\theta}/\log |I_k|^{1/2}),$$

$$\sup_{\theta \in \Theta_0} |E[h_0^{(1)}(u, \theta)] - h_0^{(1)}(u)| \leq C_{\inf} (d_{\theta}/\log |I_k|^{1/2}).$$

Following similar proof procedures of Theorems 1 and 2, Lemma 5 leads to the following theorem, which gives a regret upper bound of $O_d(\sqrt{T})$, achieving the same convergence rate with the parametric case up to logarithmic terms (Javanmard and Nazerzadeh 2019).
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 from Theorem 1 that the regret \( \text{reg} \) under the setting where the market value is bounded support and smooth cdf. Throughout this section, we focus on the high-dimensional setting.

Remark 14. Theorem 3 partly overturns the conjecture in Shah, Johari, and Blanchet (2019) that there is no policy that can achieve an \( \tilde{O}(d) \) regret under the setting where the market value is linear in the features as in (2). We provide a regime with superior smooth market noise in which \( \tilde{O}(d) \) regret upper bound is attainable by our policy.

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

5.2. Comparison with Other Methods

In this section, we provide numerical studies which illustrate the performance of our policy over "RMLP-2" and "Bandit." For each smoothness degree \( m \in \{ 2, 4, 6 \} \), we generate our data following the same synthetic and real differences between our methods and two highly related prior algorithms 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 nonparametric noise distribution that achieves \( O(\sqrt{T}) \) regret without modeling covariate information.
Figure 1. Regret log-log plot in the setting with iid 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 \( \log(\text{reg}(T)) : = \log(\text{reg}(T)) - 2 \log \log T - (\log(\text{reg}(1500))) - 2 \log \log 1500 \). The solid blue, green and red lines represent the mean \( \bar{\text{reg}}(T) \) of the Algorithm 1 with unknown \( g(\cdot) \) and \( \theta_0 \), unknown \( g(\cdot) \) but known \( \theta_0 \), and known \( g(\cdot) \) but unknown \( \theta_0 \), 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 \( \frac{1}{m} \) with \( m \in \{2, 4, 6\} \).

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 to 12,000, 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, supplementary materials) without knowing degree of smoothness \( m \). The green and red lines are the regrets of implementing “RMLP-2” and “Bandit” policy, respectively.

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, Simsek, and Ryzin 2015; Ban and Keskin 2020; Luo, Sun, and Liu 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, Sun, and Liu (2021), Wang et al. (2020a) and consider the 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, Simsek, and Ryzin (2015) and 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, Sun, and Liu (2021), and 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
bound \( \theta \) be affected by scale of the covariates). We randomly sample 5000 data points, estimate \( \theta_0 \) using semi-parametric estimation tools from these data. We next treat it as the underlying true parameter for our binary choice model stated in (3). Given these key components, the remaining experiments remain almost the same as discussed in Sections 5.1 and 5.2. We set \( B_0 = 4, \ell_0 = 200 \) and conduct Algorithm 4, supplementary materials (we use cross-validation to select \( n \) at the beginning of every episode.

We next compare Algorithm 4, supplementary materials 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 nonparametric 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 article, 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 \( F \in C^{(m)} \), we conjecture that a tighter regret lower bound \( \Omega_d(\frac{m}{\sqrt{T}}) \) 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, for example, \( v_t = \phi(\theta_0^T x_t) + z_t \) or other structured statistical machine learning model such as the additive model \( v_t = f_1(x_{1t}) + \cdots + f_d(x_{dt}) + z_t \). Finally, it’s worth studying similar pricing problems with adversarial or strategic buyers, which is potentially more suitable in some specific applications.

Supplementary Materials

The supplementary materials contain the technical proofs of theorems and relevant codes for this paper’s numerical study.

Funding

Research supported by the NSF grant DMS-2052926, DMS-2053832, and the ONR grant N00014-19-1-2120

ORCID

Mengxin Yu http://orcid.org/0000-0002-6818-4083

References

Alexey, D. (2020), "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, pp. 2668–2677. July, 13–18, PMLR. [553]

Amin, K., Rostamizadeh, A., and Syed, U. (2014), "Repeated Contextual Auctions with Strategic Buyers," in Advances in Neural Information Processing Systems (Vol. 27). [553]

Anton, Z., and Alexey, D. (2020), "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, pp. 11469–11480. PMLR. [553]

Babaioff, M., Dughmi, S., Kleinberg, R., and Slivkins, A. (2015), "Dynamic Pricing with Limited Supply," The ACM Transactions on Economics and Computation, 3, 1–26. doi: 10.1145/2559152. [553]

Ban, G., and Keskin, N. (2020), "Personalized Dynamic Pricing with Machine Learning: High Dimensional Features and Heterogeneous Elasticity," Management Science, 67, 5549–5568. [552,553,562]

Berg, A., and Politis, D. (2009), "CDF and Survival Function Estimation with Infinite-Order Kernels," Electronic Journal of Statistics, 3, 1436–1454. [560]

Besbes, O., and Zeevi, A. (2009), "Dynamic Pricing without Knowing the Demand Function: Risk Bounds and Near-Optimal Algorithms," Operations Research, 57, 1407–1420. [553]

Broder, J., and Rusmevichientong, P. (2012), "Dynamic Pricing under a General Parametric Choice Model," Operations Research, 60, 965–980. [553]

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," Journal of the American Statistical Association, 92, 477–489. [553,555]

Cesa-Bianchi, N., Cesari, T., and Perchet, V. (2019), Dynamic Pricing with Finitely Many Unknown Valuations, volume 98 of Proceedings of Machine Learning Research, pp. 247–273. PMLR. [553]

Chen, N., and Gallego, G. (2020), "Nonparametric Pricing Analytics with Customer Covariates," Operations Research, 69, 974–984. [553,554]

Chen, Q., Jasin, S., and Duenez, I. (2019), "Nonparametric Self-Adjusting Control for Joint Learning and Optimization of Multiproduct Pricing with Finite Resource Capacity," Mathematics of Operations Research, 44, 601–631. [553]

Chen, X., Simchi-Levi, D., and Wang, Y. (2021), "Privacy-Preserving Dynamic Personalized Pricing with Demand Learning," Management Science, 68, 4878–4898. [553]

Chen, X., Owen, Z., Pixton, C., and Simchi-Levi, D. (2022), "A Statistical Learning Approach to Personalization in Revenue Management," Management Science, 68, 1923–1937. [554]

Cohen, M. C., Lobel, I., and Paes Leme, R. (2016), "Feature-based Dynamic Pricing," Management Science, 66, 4921–5484. [553]

Delercoix, M., Härdle, W., and Hristache, M. (2003), "Efficient Estimation in Conditional Single-Index Regression," Journal of Multivariate Analysis, 86, 213–226. [553]

den Boer, A. V. (2015), "Dynamic Pricing and Learning: Historical Origins, Current Research, and New Directions," Surveys in Operations Research and Management Science, 20, 1–18. [552,553]
