Semi-Parametric Contextual Pricing Algorithm using Cox Proportional Hazards Model

Young-Geun Choi [1] **Gi-Soo Kim** [2] **Yunseo Choi** [3] **Wooseong Cho** [4] **Myunghee Cho Paik** [5 6] **Min-hwan Oh** [4]

Abstract

Contextual dynamic pricing is a problem of setting prices based on current contextual information and previous sales history to maximize revenue. A popular approach is to postulate a distribution of customer valuation as a function of contextual information and the baseline valuation. A semi-parametric setting, where the context effect is parametric and the baseline is nonparametric, is of growing interest due to its flexibility. A challenge is that customer valuation is almost never observable in practice and is instead *type-I interval censored* by the offered price. To address this challenge, we propose a novel semi-parametric contextual pricing algorithm for stochastic contexts, called the epoch-based Cox proportional hazards Contextual Pricing (CoxCP) algorithm. To our best knowledge, our work is the first to employ the Cox model for customer valuation. The CoxCP algorithm has a high-probability regret upper bound of $\tilde{O}(T^{\frac{2}{3}}d)$, where T is the length of horizon and d is the dimension of context. In addition, if the baseline is known, the regret bound can improve to $O(d\log T)$ under certain assumptions. We demonstrate empirically the proposed algorithm performs better than existing semi-parametric contextual pricing algorithms when the model assumptions of all algorithms are correct.

[1]Department of Mathematics Education, Sungkyunkwan University, Seoul, South Korea [2]Department of Industrial Engineering and Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea [3]Department of Statistics, Sookmyung Women's University, Seoul, South Korea [4]Graduate School of Data Science, Seoul National University, Seoul, South Korea [5]Department of Statistics, Seoul National University, Seoul, South Korea [6]Shepherd23 Inc., Seoul, South Korea. Correspondence to: Min-hwan Oh <minoh@snu.ac.kr>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

The contextual dynamic pricing problem involves setting real-time prices for products or services based on contextual factors such as product details and customer characteristics. There is a significant body of literature on this topic due to its importance and practical applications. We refer to den Boer (2015), Wei & Zhang (2018) and Mišić & Perakis (2020) for comprehensive review and Chen & Chen (2015), Hu et al. (2015), Dutta & Mitra (2017) and Saharan et al. (2020) for applications. A common goal of dynamic pricing algorithms is to maximize the seller's revenue. To achieve this, a good pricing policy should balance both learning about customer demand through exploration and setting prices based on current knowledge through exploitation.

A popular setting for contextual pricing is the customer valuation model, also known as the binary choice model, over a specific time horizon. Specifically, in each sales round $t \in [T] = \{1, \dots, T\}$, a feature $x_t \in \mathbb{R}^d$ is provided to describe the characteristics of the product and customer. The customer's valuation of the product, or market price, is represented by a random variable v_t, which is unknown to the seller. The seller proposes a price p_t based on previous sales records and the current context x_t, and then observes feedback $y_t \in \{0, 1\}$ on whether the customer purchases the product. The customer valuation model assumes that $y_t = \mathbb{1}_{v_t > p_t}$, meaning that a purchase occurs if and only if the selling price is less than the customer's valuation. As a result, the revenue at round t is $p_t y_t$, and the seller aims to maximize the expected revenue. Note that the customer valuation v_t is only partially observed, in that it is right- or left-censored by p_t. This type of data is referred to as Type-I interval censored data in the statistical literature.

Contextual pricing algorithms under the customer valuation model postulate a probabilistic model family for the conditional distribution of v_t given x_t. Let $F(p|x_t) = \mathbb{P}(v_t \leq p|x_t)$ be the cumulative distribution function (CDF) of v_t given x_t. The linear model (Javanmard & Nazerzadeh, 2019; Xu & Wang, 2021; Luo et al., 2022; Fan et al., 2022) assumes a *location family* $F(p|x_t) = F_0(p - x_t^T\beta)$, and the log-linear model (e.g., Shah et al., 2019) assumes a *scale family* $F(p|x_t) = F_0(p\exp(-x_t^T\beta))$, where $F_0(p) = \mathbb{P}(v_t \leq p|x_t = 0)$ is the baseline CDF and $\beta \in \mathbb{R}^d$ is

Table 1. Comparison of customer valuation model-based contextual dynamic pricing algorithms with stochastic contexts. Notes: ♯ Informal definition is that the baseline revenue function $p(1-F_0(p))$ is smooth at its maximizer. Exact definitions are slightly different; ♭ Informal definition is that $p(1-F_0(p))$ has a unique maximizer. Exact definitions are slightly different; § $m \geq 2$ is the order of smoothness of F_0; † s is the sparsity (number of nonzero coefficients) of the parameter vector.

METHOD	MODEL FOR v_t	REGRET UPPER BOUND	ASSUMPTIONS ON F_0: 2ND-ORDER SMOOTHNESS♯	ASSUMPTIONS ON F_0: OPTIMAL PRICE UNIQUENESS♭
(If F_0 is unknown)				
SHAH ET AL. (2019)	LOG-LINEAR	$\widetilde{O}(T^{\frac{1}{2}}d^{\frac{11}{4}})$	○	×
LUO ET AL. (2022)	LINEAR	$\widetilde{O}(T^{\frac{2}{3}}d^2)$	○	×
LUO ET AL. (2022)	LINEAR	$\widetilde{O}(T^{\frac{3}{4}}d)$	×	×
FAN ET AL. (2022)	LINEAR	$\widetilde{O}((Td)^{\frac{2m+1}{4m-1}})$§	○	○
COXCP **(This work)**	PH	$\widetilde{O}(T^{\frac{2}{3}}d)$	×	×
(If F_0 is known)				
JAVANMARD & NAZERZADEH (2019)	LINEAR	$O(s\log T)$†	○	○
XU & WANG (2021)	LINEAR	$O(d\log T)$	○	○
COXCP WITH FIXED F_0 **(This work)**	PH	$O(d\log T)$	○	○

the parameter of context effect. A popular approach is to assume F_0 is known and estimate β using the maximum likelihood principle (Javanmard & Nazerzadeh, 2019; Xu & Wang, 2021). Recently, there has been growing interest in cases where the baseline CDF F_0 is unknown and nonparametric (Shah et al., 2019; Luo et al., 2022; Fan et al., 2022), since it allows for more flexible modeling of the baseline valuation distribution. These approaches are also called semi-parametric, meaning that the seller must learn both the parametric part (β) and the nonparametric part (F_0). In general, the semi-parametric learning problem for the type-I interval censored data under (log-)linear models is very challenging. One possible reason for this difficulty is that the probability of purchase under these models has β lying inside $F_0(\cdot)$, which may hinder theoretical investigation provided that F_0 is nonparametric.

In this work, we propose a novel semi-parametric contextual dynamic pricing algorithm, namely Cox Contextual Pricing (CoxCP) algorithm, under the customer valuation model setting with stochastic contexts. The algorithm utilizes the Cox proportional hazards (PH) model. It defines a *shape family* on $F(p|x_t)$, where the baseline CDF F_0 is allowed to be nonparametric and the parametric effect $x_t^T\beta$ affects the "shape" of F_0, rather than "shifting" F_0 (as in the location family) or "stretching" F_0 (as in the scale family). The PH model has a long history, dating back to the seminal work of Cox (1972), and has been widely used to model time-to-event data in various fields including biomedical sciences, econometrics, and industrial statistics. It also has been used for any positive random variables in the presence of censoring such as healthcare cost (Malehi et al., 2015) and wage (Fortin et al., 2011). Interestingly, to the best of our understanding, the adoption of the PH model on customer valuation has not been previously studied.

The proposed CoxCP algorithm employs a nonparametric maximum likelihood estimator (NPMLE) to estimate β and F_0 of the PH model. Technically, unlike the linear and log-linear models, the probability of purchase under the PH model has β lying outside $F_0(\cdot)$, which may facilitate separate analysis of β and F_0 in terms of both theory and computation. With a judicious application of the semi-parametric estimation theory in the statistical literature related to the PH model under type-I interval censoring (e.g. Finkelstein 1986; Huang 1996; Anderson-Bergman 2017), it is feasible to establish a regret bound for the proposed algorithm over a broad range of F_0.

We remark that the distribution of customer valuation is unknown and can arise from any distributions, even with multiple modes (Wang et al., 2021b). Hence, the exploration and augmentation of the model classes that foster efficient learning are of paramount importance.

Contributions. In Table 1, we present a summary of our results and key assumptions in comparison to existing results from closely related literature. We make two theoretical contributions for the proposed algorithms.

1. We derive a high-probability regret upper bound of $\widetilde{O}(T^{\frac{2}{3}}d)$ for the CoxCP algorithm when F_0 is unknown and nonparametric, without assuming the second-order smoothness around the optimal price nor the uniqueness of the optimal price. These are common assumptions in the contextual pricing literature. Our result improves upon the existing regret bound of Luo et al. (2022), which reports $\widetilde{O}(T^{\frac{2}{3}}d^2)$ with the second-order smoothness assumption and $\widetilde{O}(T^{\frac{3}{4}}d)$ without the two

assumptions. Compared to the results of Fan et al. (2022), which reported a regret bound of $\widetilde{O}(T^{\frac{5}{7}}d^{\frac{5}{7}})$ under both conditions, our result has a lower order in T and relaxes the assumptions. Additionally, compared to the order $\widetilde{O}(T^{\frac{1}{2}}d^{\frac{11}{4}})$ of Shah et al. (2019), our algorithm has a slightly increased order in T, but it improves with respect to d and relaxes the second-order smoothness assumption.

2. For a comparison purpose, we derive an improved high-probability regret upper bound of $O(d\log T)$ for the CoxCP algorithm for the known F_0 case, provided that we additionally assume the second-order smoothness and the optimal price uniqueness. Our result and assumptions are comparable to existing regret upper bounds in linear model-based algorithms with known F_0 as reported in Javanmard & Nazerzadeh (2019); Xu & Wang (2022).

Outline of paper. In Section 2, we briefly review related literature. In Section 3, we describe the setup of the customer valuation-based dynamic pricing problem. In Section 4, we define the PH model and give its interpretation under the setup. In Section 5, we develop the proposed CoxCP algorithm. In Section 6, we derive the regret upper bound of the proposed algorithm. In Section 7, we conduct numerical study. Finally, Section 8 concludes the paper.

2. Related Works

Customer valuation model-based algorithms with stochastic contexts. A popular approach is to assume a linear customer valuation model $F(p|x_t) = F_0(p - x_t^T\beta)$ with known F_0 (Amin et al., 2014; Javanmard & Nazerzadeh, 2019; Xu & Wang, 2021). Amin et al. (2014) assumed that the market price is a deterministic linear transformation of the context, i.e., F_0 is the point mass at zero. Javanmard & Nazerzadeh (2019) assumed that both F_0 and $1-F_0$ are twice continuously differentiable and log-concave, which implies the uniqueness of the optimal price. They proposed an epoch-based regularized maximum likelihood algorithm with a regret bound of $O(s\log T)$, where s is the number of nonzero elements in the true β. Xu & Wang (2021) proposed an unregularized version of Javanmard & Nazerzadeh (2019)'s algorithm and derived the same regret bound with a relaxed assumption on $\mathbb{E}(x_t x_t^T)$.

Some pioneering works assumed F_0 as unknown and non-parametric (Shah et al., 2019; Fan et al., 2022; Luo et al., 2022). Shah et al. (2019) assumed the log-linear model $F(p|x_t) = F_0(p\exp(-x_t^T\beta))$ and constructed a non-contextual bandit-based UCB algorithm. Here, the entire space of $(\beta, F_0(\cdot))$ is discretized; each discritized point corresponds to an "arm." Under the second-order smoothness of the revenue function, the authors derived a regret upper bound $\widetilde{O}(T^{\frac{1}{2}}d^{\frac{11}{4}})$ for a theorized version of their algorithm. Fan et al. (2022) pointed out that this bound has suboptimal dependence on d. Furthermore, a major drawback of this algorithm is computational inefficiency. Its computation complexity is $O(T^{\frac{d+1}{4}})$ which grows exponentially with d, since each discretized "arm" is evaluated at least once. Therefore, the algorithm may not scale well even if d is moderately small. More computational details are discussed in Section 7 and Appendix D. On the other hand, Luo et al. (2022) and Fan et al. (2022) constructed linear model-based algorithms based on the linear model $F(p|x_t) = F_0(p-x_t^T\beta)$. Luo et al. (2022) proposed an epoch-based algorithm that alternates between a pure exploration phase and an upper confidence bound (UCB) phase. In the exploration phase, prices are drawn independently from a uniform distribution and used to estimate β by least squares. Then, in the UCB phase, the algorithm learns discretized values of $F_0(\cdot)$. Under the assumption of second-order smoothness of the expected revenue around the optimal price, they derived a regret upper bound of $\widetilde{O}(T^{\frac{2}{3}}d^2)$. Without this assumption, they derived a regret bound of $\widetilde{O}(T^{\frac{3}{4}}d)$. Fan et al. (2022) proposed another epoch-based algorithm that alternates between exploration and exploitation phases. During the exploration phase, independent uniform sampling is conducted as in Luo et al. (2022). The estimation of β is also performed in the same manner. The difference is the estimation of F_0, which is based on the Nadaraya-Watson kernel regression estimator calculated from the exploration samples. In the exploitation phase, the myopic policy with estimated parameters is used to determine p_t. The regret bound reported is $\widetilde{O}((Td)^{\frac{2m+1}{4m-1}})$, assuming that F_0 is $m \geq 2$ times continuously differentiable and the optimal price is unique. For example, if $m = 2$, the order becomes $\widetilde{O}((Td)^{\frac{5}{7}})$.

Other contextual pricing algorithms. Several researchers have proposed linear valuation model-based pricing algorithms for adversarial contexts, including Cohen et al. (2020), Liu et al. (2021), Krishnamurthy et al. (2021), and Xu & Wang (2021). All of these works assumed that F_0 is known. Golrezaei et al. (2019) proposed another linear valuation model-based contextual pricing algorithm, but with an additional assumption that side information about the buyer's bidding prices is available. Demand model-based contextual pricing algorithms, which assume that the number of sold items under the given price is the linear function of the price and contexts, are popular as well (Qiang & Bayati, 2016; Ban & Keskin, 2021; Nambiar et al., 2019; Wang et al., 2021c;a; Bu et al., 2022).

Proportional hazard (PH) models for type-I interval-censored data. Our work is related to the semi-parametric estimation problem of the PH model under type-I interval censoring. When there is no context present, Turnbull (1976) first proposed a pool-adjacent-violators algorithm-type esti-

mator for F_0, which was later clarified by Groeneboom & Wellner (1992) as the NPMLE for F_0. Finkelstein (1986) first proposed the PH model for type-I interval-censored data and presented the NPMLE of the parameters. The large-sample properties of these estimators were established by Huang (1996). For a comprehensive review of this topic, we refer the reader to Banerjee (2012); Groeneboom & Jongbloed (2014).

3. Problem Setting

We formally introduce notations and settings. We consider a stochastic contextual pricing problem under the customer valuation model. There are T consecutive sales sessions (rounds), where each round involves a single product. The overall procedure is summarized below.

For each sales round $t = 1, \dots, T$,

1. The seller observes a context vector $x_t \in \mathbb{R}^d$.
2a. The seller offers a price p_t based on x_t and the previous sales records $\{(x_\tau, p_\tau, y_\tau)\}_{\tau=1}^{t-1}$.
2b. Simultaneously, the customer evaluates the product at v_t, which is not known to the seller.
3. The seller observes $y_t = \mathbb{1}_{v_t > p_t}$, whether the product was sold or not.

It is standard to assume that x_t is independently and identically distributed (i.i.d.) and the distribution of v_t depends only on x_t. The expected revenue for any offered price p given x_t is $\mathbb{E}(p\mathbb{1}_{v_t>p}|x_t) = p\mathbb{P}(v_t > p|x_t) = p(1 - F(p|x_t))$, where we recall $F(p|x_t) = \mathbb{P}(v_t \le p|x_t)$. For concise presentation, we define the complementary CDF (reverse CDF) of v_t given x_t as $S(p|x_t) = 1 - F(p|x_t) = \mathbb{P}(v_t > p|x_t)$. The optimal price p_t^* at time t is defined by a maximizer of the expected revenue function at the round,

$$p_t^* \in \operatorname*{argmax}_p pS(p|x_t). \tag{1}$$

Note that $p_t^* = p_t^*(x_t)$ depends on x_t. The pricing policy (1) is also called the clairvoyant policy. The regret at step t is defined by the difference between the expected revenues from the optimal price p_t^* and the offered price p_t,

$$regret(t) = p_t^* S(p_t^*|x_t) - p_t S(p_t|x_t). \tag{2}$$

The goal of the seller is to devise a pricing policy that can minimize the cumulative regret $R(T) = \sum_{t=1}^{T} regret(t)$.

4. Cox Proportional Hazards (PH) Model

We propose to model the distribution of v_t by the proportion hazard (PH) model (Cox, 1972) formulated by

$$S(p|x_t) \equiv \mathbb{P}(v_t > p|x_t) = S_0(p)^{\exp(x_t^T \beta)}, \tag{3}$$

where $S_0(p) = \mathbb{P}(v_t > p|x_t = 0) = 1 - F_0(p)$ is a baseline complement CDF, and $\beta \in \mathbb{R}^d$ is a parameter vector representing the linear effect of a given context. The context effect $x_t^T \beta$ determines the shape of the distribution of v_t through the exponent $\exp(x_t^T \beta)$. If $S_0(p)$ is unknown and nonparametric, (3) is a semi-parametric model where it only assumes a specific form for the context effect. If $S_0(p)$ is known, (3) becomes a classic parametric model.

Intuition. For thorough comprehension, we provide interpretation of the PH model within the customer valuation setting. Let $F(p)$, $S(p) = 1 - F(p)$ and $f(p) = F'(p)$ be the CDF, complementary CDF, and density function of v_t. The "hazard rate"[1] function of v_t is defined by

$$\lambda(p) := \frac{f(p)}{S(p)} = -\frac{d}{dp}\{\log S(p)\}. \tag{4}$$

Similarly, we denote by $f(p|x_t)$ and $\lambda(p|x_t)$ the conditional density and hazard function of v_t given x_t. By definition, for a small $\Delta p > 0$, $\lambda(p)\Delta p = f(p)\Delta p/\mathbb{P}(v_t > p)$ approximates the conditional proportion of customers who value a product between p and $p + \Delta p$, given that they would purchase it if the price were not to exceed p. As these customers would decline to buy the product if the price increases to $p + \Delta p$, $\lambda(p)\Delta p$ can also be interpreted as the approximate conditional proportion of customer attrition, or churn, resulting from a minor price increase. Another perspective on $\lambda(p)$ is to consider it as the negative Rate of Return (RoR) of demand at price p. This interpretation is backed by the fact that $\lambda(p) = -S'(p)/S(p)$ and $S(p)$ can be perceived as the demand at price p. An additional remark is that if p represents a logarithmic price, then $\lambda(p)$ aligns with the price elasticity at price p.

The PH model postulates that the hazard function of a product, given its context, is proportional to the baseline hazard function. From (4), one can easily check that the PH model (3) can be restated as

$$\lambda(p|x_t) = \lambda_0(p) \exp(x_t^T \beta), \tag{5}$$

where $\lambda_0(p) := F_0'(p)/S_0(p)$ is the baseline hazard rate function. Essentially, the PH model presumes that the RoR of demand functions is proportional across different contexts. For instance, if a given covariate is binary with $d = 1$, $\exp(\beta) = \lambda(p|x_t = 1)/\lambda(p|x_t^* = 0)$ signifies the ratio of the RoR of demand at a session with $x_t = 1$ to that with $x_t = 0$. It's noteworthy that the PH model is semi-parametric, allowing the assumptions on the baseline $\lambda_0(\cdot)$ to remain minimal. In Appendix A, we provide a toy example of a customer valuation equipped with the PH model to facilitate better understanding.

[1] The term originates from the survival analysis.

Optimal price revisited. Under (3), the definition of the optimal price p_t^* (1) is rewritten as

$$p_t^* \in \operatorname*{argmax}_p pS(p|x_t) = \operatorname*{argmax}_p pS_0(p)^{\exp(x_t^T\beta)}. \quad (6)$$

Since the context effect is acting on the exponent of $S_0(\cdot)$, the support of v_t remains the same with the support of $F_0(\cdot)$. As p_t^* lies on the support of v_t, we can conclude that p_t^* always lies on the support of F_0 regardless of the values of given x_t. This property is unique in the PH model compared to the (log)-linear models. The linear model is a location shift family where the parameter part $x_t^T\beta$ determines the amount of translation of F_0, and the log-linear model is a scale shift where $\exp(x_t^T\beta)$ determines the scale. Thus, the support of v_t shifts according to $x_t^T\beta$ in both models. If an outlying value of x_t is input, then corresponding p_t^* may be an unreasonable, for example, negative or extremely large price. On the other hand, in the PH model framework, we can anticipate that the optimal prices will lie on the baseline support even for accidentally large x_t.

Connection to the log-linear model. In the statistical literature, the PH model and the log-linear model (also known as the accelerated failure time model) are the most common choices to account for contextual effects in the survival analysis. Although they assume different families of distributions (shape family and scale family), a special case of the log-linear model is included in the PH model. To illustrate this, we recall that an equivalent formulation on the log-linear model $S(p) = S_0(p\exp(-x_t^T\beta))$ is $\log v_t = x_t^T\beta + \epsilon_t$, where S_0 is the complement CDF of $\exp(\epsilon_t)$. When the distribution of ϵ_t is extreme value distribution indexed by scale parameter σ, so that $\epsilon_t = \sigma\epsilon_t^*$ with ϵ_t^* arising from standard extreme value distribution, the model also belongs to the PH model with $\lambda(p|x_t) = \lambda_0(p)\exp(x_t^T\beta)$, where $\lambda_0(t) = \alpha t^{\alpha-1}$.

5. Proposed Algorithm

Epoch-based design. We employ an epoch-based design (also known as the doubling trick) that segments the given horizon T into several clusters of rounds ("epochs") and executes identical pricing policies on a per-epoch basis. In every k-th epoch, the offered price is constructed from (6), where the true β and S_0 are estimated from the data at the previous epoch. Such design has been employed in maximum likelihood estimator-based pricing algorithms (Javanmard & Nazerzadeh, 2019; Xu & Wang, 2021). A complete pseudocode is provided in Algorithm 1. In the following portion of this section, we examine the details of parameter estimation and algorithmic variations of the proposed algorithm.

Algorithm 1 Cox Contextual Pricing (CoxCP) algorithm

Input: The length of the first epoch, τ_1; the minimum and maximum of price search range, $p_{\min}$ and $p_{\max}$.
For $t = 1, \dots, \tau_1$, observe x_t, randomly choose p_t from a distribution supported on $[p_{\min}, p_{\max}]$, and get reward y_t;
for epoch $k = 2, 3, \dots$ **do**
  Estimate $\theta = (\beta, S_0(\cdot))$ by the NPMLE,

$$\widehat{\theta}^k \leftarrow \operatorname*{argmin}_\theta L_{k-1}(\theta),$$

  where $L_{k-1}(\theta)$ is defined in (9);
  Set $\tau_k \leftarrow 2\tau_{k-1}$;
  Set $\mathcal{E}_k \leftarrow \{\sum_{r=1}^{k-1}\tau_r + 1, \dots, \sum_{r=1}^{k}\tau_r\}$;
  for round $t \in \mathcal{E}_k$ **do**
    Observe x_t;
    Offer price by the myopic policy,

$$p_t \in \operatorname*{argmax}_{p\in[p_{\min},p_{\max}]} \left\{p\widehat{S}_0^k(p)^{\exp(x_t^T\widehat{\beta}^k)}\right\}; \quad (7)$$

    Get reward y_t.
  end for
end for

Nonparametric maximum likelihood estimation of parameters. Let $\mathcal{E}_k$ denote the set of round indices for epoch k. Our algorithm design ensures that, conditioned on the data from previous epochs $1, \dots, k-1$, the data for each epoch k, $\{(x_t, p_t, y_t)\}_{t\in\mathcal{E}_k}$, is independently and identically distributed (i.i.d.) over t. Furthermore, the algorithm's configuration guarantees that $v_t \perp p_t|x_t$ for each $t \in \mathcal{E}_k$. This is due to the fact that p_t solely relies on x_t and previous epochs, while the distribution of v_t only depends on x_t. Consequently, if we employ Huang (1996)'s Nonparametric Maximum Likelihood Estimator (NPMLE) using $\{(x_t, p_t, y_t)\}_{t\in\mathcal{E}_k}$, it can be shown that it consistently estimates the true parameter $(\beta, S_0(\cdot))$ of the PH model (3), subject to certain mild assumptions.

For completeness, we describe key estimation steps.

Let $\theta = (\beta, S_0(\cdot))$ be the estimation target. Since $y_t|(p_t, x_t)$ is a binary random variable with success probability $S(p_t|x_t)$, a negative loglikelihood function of θ given tuple (x_t, p_t, y_t) at round t is

$$\begin{aligned} l_t(\theta) &= -y_t\log\{S(p_t|x_t)\} - (1-y_t)\log\{F(p_t|x_t)\}, \\ &= -y_t\exp(x_t^T\beta)\log\{S_0(p_t)\} \\ &\qquad -(1-y_t)\log\{1 - S_0(p_t)^{\exp(x_t^T\beta)}\}, \end{aligned} \quad (8)$$

up to an additive constant not depending on θ. Then, we can define the negative loglikelihood of θ given the data at epoch k as $L_k(\theta) = \sum_{t\in\mathcal{E}_k} l_t(\theta)$. For the convenience of

discussion, we further define the baseline cumulative hazard function as $\Lambda_0(p) := \int_0^p \lambda_0(v)dv$. From (4), it holds that $\Lambda_0(p) = -\log S_0(p)$. Since $S_0(\cdot)$ and $\Lambda_0(\cdot)$ is one-to-one, we can reparametrize θ as $\theta = (\beta, \Lambda_0(\cdot))$, with a slight notational abuse. Then $L_k(\theta)$ is

$$L_k(\theta) = \sum_{t\in\mathcal{E}_k} \Big\{ y_t \Lambda_0(p_t) \exp(x_t^T\beta) - (1-y_t)\log\{1-\exp(-\Lambda_0(p_t)\exp(x_t^T\beta)\}\Big\}. \quad (9)$$

A benefit of the reparametrization by Λ_0 instead of S_0 is that the $L_k(\theta)$ is convex in each $\Lambda_0(p_t)$ (if viewed as a scalar parameter). The estimation procedure minimizes $L_k(\theta)$ as a function of β and $\{\Lambda_0(p_t)\}_{t\in\mathcal{E}_k}$, so $d+|\mathcal{E}_k|$ "parameters" in total, under the constraint that $\Lambda_0(\cdot)$ is monotone increasing (i.e, $\Lambda_0(p_{t_1}) \le \Lambda_0(p_{t_2})$ for any $p_{t_1} \le p_{t_2}$). The MLE of β is the corresponding solution of the minimization problem. The NPMLE of Λ_0 is a right-continuous step function in which jumps occur only at $\{p_t\}_{t\in\mathcal{E}_k}$ and the function value at each p_t is equal to the corresponding solution for $\Lambda_0(p_t)$. Then, the NPMLE of S_0 is $\widehat{S}_0(\cdot) = \exp(-\widehat{\Lambda}_0(\cdot))$. Finally, we denote the (NP)MLE of θ by $\widehat{\theta}^{k+1}$ to emphasize that the data in epoch k is used in offering price in epoch $k+1$. Note that (9) is a constrained convex minimization problem that can be efficiently solved with convergence guarantees. For example, Finkelstein (1986), Huang (1996) and Anderson-Bergman (2016) proposed algorithms based on an alternating optimization of Λ_0 and β. In particular, Anderson-Bergman (2016)'s algorithm is available in `R` package `icenReg` (Anderson-Bergman, 2017), which we employed to implement our algorithm in Section 7.

Price offering. Let $k \ge 2$. For each round t in epoch k, we set the offered price p_t by (7), which is a myopic version of (6). We propose a grid search for solving (7) since it is a one-dimensional maximization problem on a closed interval. For rounds t in the first epoch $\mathcal{E}_1$, as an initial step, we offer randomly sampled prices.

We note that a well-known, randomized ϵ-greedy heuristic (Auer et al., 2002) can be injected at the stage of price offering (7) to encourage exploration. To be specific, let $\alpha_k \in [0,1]$ be a constant depending on epoch k. For each round $t \in \mathcal{E}_k$, the ϵ-greedy heuristic offers p_t by (7) with probability $1-\alpha_k$, and otherwise chooses a random price. It is easy to check that choosing $\alpha_k = \min\{\gamma 2^{-\frac{k-1}{3}}, 1\}$ does not change the regret upper bound of Algorithm 1 (Theorem 6.1 in Section 6), where γ is a global constant. In our simulation experiment, we tune γ to optimize the degree of exploration. For completeness, we provide the full pseudocode of the CoxCP algorithm adding the ϵ-greedy heuristic in Algorithm 2 of the Appendix C.

Our theoretical assumptions suggest that, whenever the random sampling of p_t is needed, it suffices to sample from any arbitrary distribution supported on $[p_{\min}, p_{\max}]$, where the technical assumptions on $p_{\min}$ and $p_{\max}$ will be provided later. One may want to set $p_{\min}$ as zero and $p_{\max}$ as a large value if she has no information on the prices. If some prior information exists, then one can inject prior knowledge into the sampling distribution of p_t. We remark that it is not necessary to uniformly sample from $[p_{\min}, p_{\max}]$. On the other hand, the linear model-based algorithms (Fan et al., 2022; Luo et al., 2022) require sampling from a specific distribution (the uniform distribution) to guarantee the consistency of β.

Modified algorithm when F_0 is known. When the true baseline valuation F_0 is known, one can plug-in the true S_0 (equivalently Λ_0) to the likelihood (9) and the myopic policy (7). In the next Section, we prove that the resulting algorithm has improved regret bound compared to the original Algorithm 1 which matches the bound from the myopic policy-based algorithms proposed under the linear model (Javanmard & Nazerzadeh, 2019; Xu & Wang, 2021). If one considers to inject the ϵ-greedy heuristic, it is easy to check $\alpha_k = \min\{\gamma 2^{-(k-1)}, 1\}$ does not change the regret upper bound (Theorem 6.5 of Section 6).

6. Regret Analysis

We present our main theorems and proofs. To save space, we will defer some details to the Appendices.

6.1. Assumptions

First, we make the following assumptions:

Assumption 1 (Bounded parameter space)**.** True β lies in the interior of $\{\beta \in \mathbb{R}^d : \|\beta\|_2 \le \mathcal{B}\}$ for some $\mathcal{B} > 0$.

Assumption 2 (Bounded i.i.d. contexts)**.** (a) $x_t \in \mathbb{R}^d$ is independently and identically (i.i.d.) drawn a distribution that does not involve β and F_0, and $\|x_t\|_2 \le \mathcal{X}$ with probability 1 for some $\mathcal{X} > 0$. (b) For any $\beta_1 \ne \beta_2$, $\mathbb{P}(x_t^T\beta_1 \ne x_t^T\beta_2) > 0$.

Assumption 3 (Mildness of baseline CDF over the search range)**.** (a) $F_0(0) = 0$; (b) F_0 has strictly positive and continuous derivative over interval $(p_{\min}, p_{\max})$; (c) $M_1 < F_0(p_{\min})$ and $F_0(p_{\max}) < M_2$ for some $0 < M_1 < M_2 < 1$; (d) For any (x, β) with $\|x\|_2 \le \mathcal{X}$ and $\|\beta\|_2 \le \mathcal{B}$, at least one maximizer of $pS_0(p)^{\exp(x^T\beta)}$ over $[0,\infty)$ lies inside $(p_{\min}, p_{\max})$.

Assumptions 1, 2a, are common in the dynamic pricing literature (Javanmard & Nazerzadeh, 2019; Shah et al., 2019; Xu & Wang, 2021; Luo et al., 2022; Fan et al., 2022). Our algorithm does not require knowing $\mathcal{B}$ and $\mathcal{X}$ in advance. Assumption 2b guarantees the model identifiability in the PH model and is equivalent to the full-rank condition of $\mathbb{E}(x_t x_t^T)$. Assumption 3 is similar to but slightly weaker

than those in the literature, in that the assumption is imposed only an interval inside F_0. We remark that, unlike existing works with unknown F_0 in the (log)-linear models (Shah et al., 2019; Fan et al., 2022; Luo et al., 2022), we allow the support of F_0 to extend beyond a bounded interval. Instead, we assume that the pre-specified search range $[p_{\min}, p_{\max}]$ contains an optimal price as in Assumption 3d.

6.2. Regret Upper Bound for CoxCP Algorithm

We provide our main result, the high-probability regret upper bound for Algorithm 1 with unknown nonparametric F_0.

Theorem 6.1. *Suppose that Assumptions 1-3 hold. For any small $\delta > 0$ and sufficiently large T, there exists $C > 0$ such that, the cumulative regret of Algorithm 1 is*

$$R(T) \leq CT^{\frac{2}{3}}d \tag{10}$$

with probability at least $1 - \delta$.

Compared with existing works with unknown baseline valuation (Shah et al., 2019; Fan et al., 2022; Luo et al., 2022), we do not make assumptions related to the second- or higher-order smoothness on $pS_0(p)$. Furthermore, compared with Fan et al. (2022) we do not require the price uniqueness assumption. We provide a sketch of the proof below and a complete proof for the Theroem and Lemmas in Appendix B.1.

Proof sketch. Fix $k \geq 2$ and let $t \in \mathcal{E}_k$ be a round in epoch k. Let $\widehat{S}^k(p|x_t) = \widehat{S}_0^k(p)^{\exp(x_t^T\beta)}$. By the definition of p_t in (7), $p_t^*\widehat{S}^k(p_t^*|x_t) - p_t\widehat{S}^k(p_t|x_t) \leq 0$. Combining this with Assumption 3d, the regret at round t is decomposed and upper bounded by

$$\begin{aligned} &regret(t) = p_t^*S(p_t^*|x_t) - p_tS(p_t|x_t) \\ &= \left\{p_t^*S(p_t^*|x_t) - p_t^*\widehat{S}^k(p_t^*|x_t)\right\} + \left\{p_t^*\widehat{S}^k(p_t^*|x_t)\right. \\ &\qquad \left. -p_t\widehat{S}^k(p_t|x_t)\right\} + \left\{p_t\widehat{S}^k(p_t|x_t) - p_tS(p_t|x_t)\right\} \\ &\leq |R_{k,t}(p_t^*)| + 0 + |R_{k,t}(p_t)|, \end{aligned} \tag{11}$$

where $R_{t,k}(u) := u\widehat{S}^k(u|x_t) - uS(u|x_t)$. To further decompose (11), we introduce the following lemma first.

Lemma 6.2. *Under the PH model (3) and Assumptions 1-3 for any small $\delta > 0$, there exist $C, K_0 > 0$ such that three terms $\|\widehat{\beta}^k - \beta\|_2$, $\|\widehat{S}_0^k(\cdot) - S_0(\cdot)\|_{L_2(Q_k)}$ and $\|\widehat{S}_0^k(\cdot) - S_0(\cdot)\|_{L_2(Q_k^*)}$ are less than $C|\mathcal{E}_k|^{-\frac{1}{3}}d$ with probability at least $1 - \delta$, for all $k \geq K_0$. Here, Q_k and Q_k^* are the marginal probability measures of p_t and p_t^* at epoch k, and $\|f(\cdot)\|_{L_2(Q)} := \{\int f(u)^2 dQ(u)\}^{\frac{1}{2}}$ for a real-valued function f and a Borel measure Q.*

Lemma 6.2 is a reexamination and modification of Theorem 3.3 of Huang (1996). Compared to the original statement, we unveiled terms related to d in the bound $C|\mathcal{E}_k|^{-\frac{1}{3}}d$. In addition, we applied our assumptions and design to derive the $L_2(Q_k)$- and $L_2(Q_k^*)$-convergence of $\widehat{S}_0^k$. The parameter estimation consistencies by Lemma 6.2 leads to the following bound:

Lemma 6.3. *Let Assumptions 1-3 hold. For a small $\delta > 0$, there exist constants $K_0, C_1, C_2 > 0$ such that for any $u \in [p_{\min}, p_{\max}]$ and $t \in \mathcal{E}_k$,*

$$|R_{t,k}(u)| \leq C_1|\widehat{S}_0^k(u) - S_0(u)| + C_2\|\widehat{\beta}^k - \beta\|_2$$

with probability at least $1 - \delta$, for all $k > K_0$.

Then, summing up the regret for all $t \in \mathcal{E}_k$, (11) yields

$$\sum_{t\in\mathcal{E}_k} regret(t) \leq 2C_2J_{1k} + C_1J_{2k} + C_1J_{3k} \tag{12}$$

whenever the event in Lemma 6.3 is true, where $J_{1k} = |\mathcal{E}_k|\|\widehat{\beta}^k - \beta\|_2$, $J_{2k} = \sum_{t\in\mathcal{E}_k}|\widehat{S}_0^k(p_t) - S_0(p_t)|$ and $J_{3k} = \sum_{t\in\mathcal{E}_k}|\widehat{S}_0^k(p_t^*) - S_0(p_t^*)|$. By Lemma 6.2, the order of J_{1k} is $|\mathcal{E}_k|^{\frac{2}{3}}d$. For J_{2k} and J_{3k}, since $J_{2k}/|\mathcal{E}_k|$ and $J_{3k}/|\mathcal{E}_k|$ are empirical approximations of $\int|\widehat{S}_0^k(u) - S_0(u)|dQ_k(u)$ and $\int|\widehat{S}_0^k(u) - S_0(u)|dQ_k^*(u)$, a combination of Lemma 6.2, the central limit theorem and Jensen's inequality yields

Lemma 6.4. *If Assumptions 1-3 hold, for small $\delta > 0$, there exists $C, K_0 > 0$ such that J_{2k} and J_{3k} are less than $C|\mathcal{E}_k|^{\frac{2}{3}}d$ with probability at least $1 - \delta$ for all $k > K_0$.*

By the Lemma above, the right-hand side of (12) has leading term $|\mathcal{E}_k|^{\frac{2}{3}}d$. Finally, summing up this regret up to epochs $k = 1$ through $\lceil\log(T/\tau_1) + 1\rceil$ concludes the proof. □

Discussion on lower bound. The minimax lower bound for the PH model-based contextual pricing problem is at least $\Omega(T^{\frac{2}{3}})$, when the true distribution class satisfies Assumptions 1–3. To show this, we employ the subclass of CDFs constructed in equation (10) of Xu & Wang (2022). This class is a subclass of the PH model family and satisfies the aforementioned assumptions. Since Xu & Wang (2022) derived an $\Omega(T^{\frac{2}{3}})$ noncontextual minimax lower bound for this class, it follows that our PH model family maintains the same lower bound. Thus, our regret upper bound $O(T^{\frac{2}{3}}d)$ matches the lower bound in terms of T. Nevertheless, a gap of order d persists; its exploration is earmarked for future research.

6.3. Regret Upper Bound for CoxCP Algorithm when F_0 is Known

Javanmard & Nazerzadeh (2019) and Xu & Wang (2021) established an improved regret bound of $O(d\log T)$ for the linear valuation model when F_0 is known. We claim that this result goes parallel in the PH model as well; we derive a logarithmic regret upper bound for the CoxCP algorithm with known F_0.

We first present additional assumptions.

Assumption 4 (Second-order smoothness of F_0)**.** F_0 is twice continuously differentiable on an interval containing $[p_{\min}, p_{\max}]$.

Assumption 5 (Optimal price uniqueness)**.** There exists $c_\psi > 0$ such that a map $\psi : p \mapsto p\lambda_0(p)$ satisfies $\psi'(p) \geq c_\psi$ for all $p \in [p_{\min}, p_{\max}]$.

Assumption 4 guarantees the regret at each round t can be bounded by $(p_t^* - p_t)^2$ up to a constant by the second-order Taylor expansion. It matches Javanmard & Nazerzadeh (2019), Xu & Wang (2021) and Fan et al. (2022)'s assumptions (when $m = 2$ in their notation). In fact, Assumption 4 can easily be weakened to the condition that $pS_0(p)^v$ can have a second-order polynomial approximation from the optimal price for any $v > 0$. Similar weaker assumptions were made in Shah et al. (2019) and Luo et al. (2022). Assumption 5 is an analogy of Assumption 2.1 of Fan et al. (2022), which guarantees that the optimal price is unique in the setting of the PH model. A sufficient condition is that $S_0(p)$ is log-concave as in Javanmard & Nazerzadeh (2019) and Xu & Wang (2021), then $c_\psi \geq \lambda_0(p_{\min})$. Another sufficient condition for Assumption 5 is the monotone hazard ratio assumption that $\lambda_0'(\cdot) \geq c_{\lambda_0}$ for a constant $C_{\lambda_0} > 0$. Additional assumptions lead to the following theorem.

Theorem 6.5. *Suppose that Assumptions 1-5 hold. For any small $\delta > 0$ and sufficiently large T, there exists $C > 0$ such that the cumulative regret of Algorithm 1 assuming F_0 as known is*

$$R(T) \leq Cd \log T \tag{13}$$

with probability at least $1 - \delta$.

Proof. Key differences from the previous subsection are the characterization of optimal prices and the decomposition of regret; for other arguments, we follow the proof of Theorem 6.1. Note that p_t^* is a maximizer of $\log(pS(p))$. Combining with $\Lambda_0(p) = -\log S_0(p)$, the first-order condition implies

$$(p_t^*)^{-1} - \exp(x_t^T\beta)\lambda_0(p_t^*) = 0,$$

equivalently $p_t^*\lambda_0(p_t^*) = \exp(-x_t^T\beta)$. By Assumption 5, a map $\psi(p) : p \mapsto p\lambda_0(p)$ is injective on $[p_{\min}, p_{\max}]$. Thus,

$$p_t^* = \psi^{-1}(\exp(-x_t^T\beta)),$$

and similarly p_t from (7) in the Algorithm 1 (with F_0 fixed) is characterized as $p_t = \psi^{-1}(\exp(-x_t^T\widehat{\beta}^k))$ for a round $t \in \mathcal{E}_k$. Combining with Assumption 4,

$$\begin{aligned} & regret(t) \leq D_1(p_t^* - p_t)^2 \\ = \;& D_1[\psi^{-1}(\exp(-x_t^T\beta)) - \psi^{-1}(\exp(-x_t^T\widehat{\beta}^k))]^2 \\ \leq \;& D_1D_2[x_t(\widehat{\beta}^k - \beta)]^2 \leq D_1D_2\mathcal{X}^2\|\widehat{\beta}^k - \beta\|_2^2, \end{aligned} \tag{14}$$

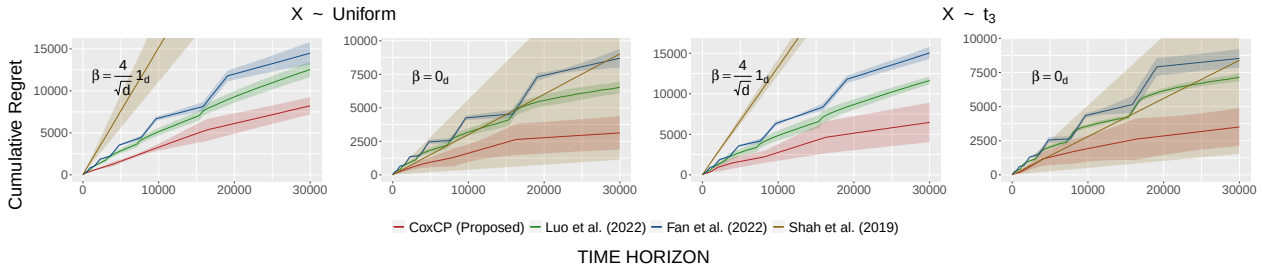
where the second and the third inequalities hold from the Mean Value theorem and the Cauchhy-Schwartz inequality, respectively, and D_1 and D_2 are global constants. Since F_0 is known, the $\widehat{\beta}^k$ is a maximum likelihood estimator for a parametric model. Thus, a standard likelihood theory implies that $\|\widehat{\beta}^k - \beta\|_2$ has order $\sqrt{d/|\mathcal{E}_k|}$. Combining with (14), we can upper bound $regret(t)$ with the leading order $d/|\mathcal{E}_k|$, which implies that the total regret at epoch k is of order d. Summing this regret for epoch $k = 1$ through $\lceil \log(T/\tau_1) + 1 \rceil$ leads to the desired result. □

Remark 6.6. The assumptions 4 and 5 are crucial to guarantee the $\log T$ order in the regret upper bound. Without those assumptions, it is straightforward to derive the order of $\sqrt{T}$.

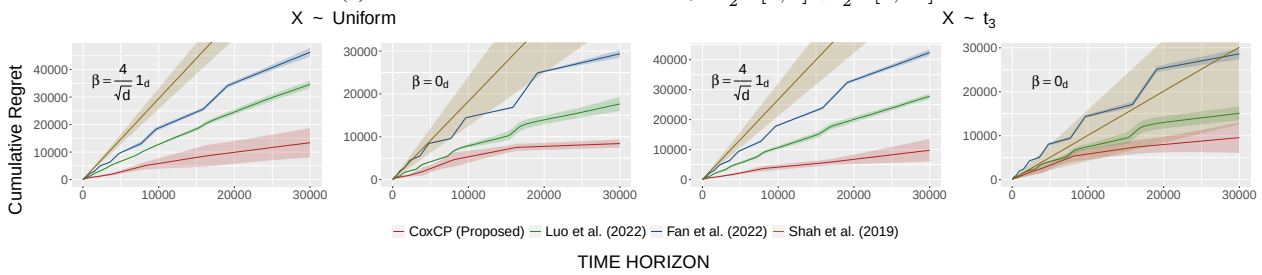
7. Simulation Experiments

We evaluate the performance of the CoxCP algorithm through Monte Carlo simulation experiments. For comparison, we included other semi-parametric algorithms for unknown baseline valuation (Shah et al., 2019; Luo et al., 2022; Fan et al., 2022). While these algorithms assume different models on $F(p|x_t)$, such a comparative study in the unknown baseline setting has yet been explored in the literature.

The total horizon was set to $T = 30{,}000$. We generated v_t following the PH model (3) with the dimension of context as $d = 5$. For true $\beta \in \mathbb{R}^d$, we considered $\beta = \frac{4}{\sqrt{d}}1_d$ and $\beta = 0_d$, where 1_d and 0_d are d-dimensional vectors of ones and zeros, respectively. In the latter scenario, all the algorithms operate under the correct model, which enables a fair comparison of their performances. Conversely, in the former scenario, we can explore the benefits when only our algorithm adheres to the correct model assumption. For sampling distributions of $x_t \in \mathbb{R}^d$, we considered a uniform distribution on d-dimensional ball with radius $\frac{1}{2}$, and multivariate t-distribution with the degree of freedom as 3 and the scale parameter as $\frac{1}{4\cdot 3(d+2)}I_{d\times d}$, where $I_{d\times d}$ is the $d \times d$ identity matrix. We intended to equalize the covariance matrices of the two distributions. As for true baseline valuation F_0, we considered two mixture distributions: $F_0 = \frac{1}{2}\mathrm{U}[1,4] + \frac{1}{2}\mathrm{U}[4,10]$ and $F_0 = \frac{3}{4}\mathrm{TN}(3.25, 0.5^2, 1, 10) + \frac{1}{4}\mathrm{TN}(7.75, 0.5^2, 1, 10)$, where $\mathrm{TN}(\mu, \sigma^2, a, b)$ is the truncated normal distribution with support $[a, b]$, location parameter μ and scale parameter σ^2. The combination of choices of true β, the distribution of x_t, and F_0 leads to eight scenarios in total. We note that every dynamic pricing algorithm involves a set of hyperparameters to control the degree of exploration. For example, the exploration of the CoxCP algorithm can be controlled through the first-epoch length τ_1 and the forced sampling frequency $\alpha_k = \min\{\gamma 2^{-(k-1)/3}, 1\}$. We conducted a grid search for $t_0 = 3{,}000$ rounds, with $\tau_1 \in \{64, 128, 256, 512, 1024\}$



(a) When the baseline distribution F_0 is $\frac{1}{2}\mathrm{U}[1,4] + \frac{1}{2}\mathrm{U}[4,10]$.



(b) When the baseline distribution F_0 is $\frac{3}{4}\mathrm{TN}(3.25, 0.5^2, 1, 10) + \frac{1}{4}\mathrm{TN}(7.75, 0.5^2, 1, 10)$.

Figure 1. Cumulative regret curves for the proposed algorithm, compared with Shah et al. (2019), Luo et al. (2022), and Fan et al. (2022)'s algorithms. Averages and standard errors over replications are marked as solid lines and bands.

and $\gamma \in \{2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{-0}\}$. The best hyperparameter was identified in the sense of the cumulative realized revenue $\sum_{t=1}^{t_0} y_t p_t$. We continued the algorithm with the best hyperparameter for the remaining $T - t_0$ rounds. For a fair comparison, we tuned the other algorithms' hyperparameters as well. In Appendix D.1, we provide more configuration details and a link to our simulation codes.

Figure 1 reports the cumulative regrets of the algorithms against rounds over five replications. When $\beta = \frac{4}{\sqrt{d}} 1_d$ where only our algorithm assumes the correct model, our algorithm achieved the best performance with a large margin, as expected. Interestingly, when $\beta = 0_d$ where all the algorithms' model assumptions are correct, our algorithm still recorded the lowest cumulative regret on average. This observation was consistent across F_0s and x_t-distributions we considered. We hypothesize that the CoxCP algorithm might have learned F_0 and β better than other algorithms. However, further empirical research is needed to better understand algorithms' behaviors in various scenarios. In addition, we reported computation times of all the algorithms in Table 2 of Appendix D.2. The computation time of our algorithm was approximately 20 seconds for running the entire $T = 30{,}000$ horizon, which was similar to Luo et al. (2022) and ∼500x faster than Fan et al. (2022) and ∼1000x faster than Shah et al. (2019).

8. Concluding Remarks

This paper proposes CoxCP, a novel contextual dynamic pricing algorithm. Our algorithm assumed the Cox proportional hazards model, a new model family, on customer valuation. The algorithm is semi-parametric and may be well suited to practical problems where customer valuation is rarely observable. The CoxCP algorithm improves on the existing algorithms' theoretical results with fewer assumptions. Our simulation study demonstrates the advantage of the proposed algorithm.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. 2020R1G1A1A01006229, 2020R1A2C1A01011950, and 2022R1C1C1006859, RS-2023-00252026) and the Institute of Information & communications Technology Planning & evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2020-0-01336, Artificial Intelligence Graduate School Program (UNIST); No. 2022-0-00469, Development of Core Technologies for Task-oriented Reinforcement Learning for Commercialization of Autonomous Drones).

References

Amin, K., Rostamizadeh, A., and Syed, U. Repeated Contextual Auctions with Strategic Buyers. *Advances in Neural Information Processing Systems (NeurIPS) 2014*,

27, 2014.

Anderson-Bergman, C. Revisiting the Iterative Convex Minorant Algorithm for Interval Censored Survival Regression Models. *Preprint*, 2016.

Anderson-Bergman, C. `icenReg`: Regression Models for Interval Censored Data in R. *Journal of Statistical Software*, 81(12), 2017.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47:235–256, 2002.

Ban, G. and Keskin, N. B. Personalized Dynamic Pricing with Machine Learning: High-Dimensional Features and Heterogeneous Elasticity. *Management Science*, 67(9): 5549–5568, 2021.

Banerjee, M. Current Status Data in the Twenty-First Century: Some Interesting Developments. In *Interval-Censored Time-to-Event Data*, 45–90. CRC Press, 2012.

Bu, J., Simchi-Levi, D., and Wang, C. Context-Based Dynamic Pricing with Separable Demand Models. *SSRN Electronic Journal* No. 4140550, 2022.

Chen, M. and Chen, Z.-L. Recent Developments in Dynamic Pricing Research: Multiple Products, Competition, and Limited Demand Information. *Production and Operations Management*, 24(5):704–731, 2015.

Cohen, M. C., Lobel, I., and Paes Leme, R. Feature-Based Dynamic Pricing. *Management Science*, 66(11):4921–4943, 2020.

Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

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

Dutta, G. and Mitra, K. A Literature Review on Dynamic Pricing of Electricity. *Journal of the Operational Research Society*, 68(10):1131–1145, 2017.

Fan, J., Guo, Y., and Yu, M. Policy Optimization Using Semiparametric Models for Dynamic Pricing. To appear at *Journal of the American Statistical Association*, Published online 2022.

Finkelstein, D. M. A Proportional Hazards Model for Interval-Censored Failure Time Data. *Biometrics*, 42 (4):845–854, 1986.

Fortin, N., Lemieux, T., and Firpo, S. Decomposition Methods in Economics. In *Handbook of Labor Economics*, 4:1–102. Elsevier Inc., 2011.

Golrezaei, N., Jaillet, P., and Liang, J. C. N. Incentive-aware Contextual Pricing with Non-parametric Market Noise. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR 206:9331–9361, 2023.

Groeneboom, P. and Jongbloed, G. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, Cambridge, 2014.

Groeneboom, P. and Wellner, J. A. *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser Basel, Basel, 1992.

Hu, Z., Kim, J.-h., Wang, J., and Byrne, J. Review of Dynamic Pricing Programs in the U.S. and Europe: Status Quo and Policy Recommendations. *Renewable and Sustainable Energy Reviews*, 42:743–751, 2015.

Huang, J. Efficient Estimation for the Proportional Hazards Model with Interval Censoring. *Annals of Statistics*, 24 (2):540–568, 1996.

Javanmard, A. and Nazerzadeh, H. Dynamic Pricing in High-dimensions. *Journal of Machine Learning Research*, 20:1–49, 2019.

Krishnamurthy, A., Lykouris, T., Podimata, C., and Schapire, R. Contextual Search in the Presence of Irrational Agents. *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC) 2021*, 53:910–918, 2021.

Liu, A., Leme, R. P., and Schneider, J. Optimal Contextual Pricing and Extensions. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA) 2021*, 1059–1078, 2021.

Luo, Y., Sun, W. W., and Liu, Y. Contextual Dynamic Pricing with Unknown Noise : Explore-then-UCB Strategy and Improved Regrets. *Advances in Neural Information Processing Systems (NeurIPS) 2022*, 36, 2022.

Malehi, A. S., Pourmotahari, F., and Angali, K. A. Statistical Models for the Analysis of Skewed Healthcare Cost Data: a Simulation Study. *Health Economics Review*, 5(1), 2015.

Mišić, V. V. and Perakis, G. Data Analytics in Operations Management: A Review. *Manufacturing & Service Operations Management*, 22(1):158–169, 2020.

Nambiar, M., Simchi-Levi, D., and Wang, H. Dynamic Learning and Pricing with Model Misspecification. *Management Science*, 65(11):4980–5000, 2019.

Qiang, S. and Bayati, M. Dynamic Pricing with Demand Covariates. *SSRN Electronic Journal* No. 2765257, 2016.

Saharan, S., Bawa, S., and Kumar, N. Dynamic pricing techniques for Intelligent Transportation System in smart cities: A systematic review. *Computer Communications*, 150:603–625, 2020.

Shah, V., Blanchet, J., and Johari, R. Semi-Parametric Dynamic Contextual Pricing. *Advances in Neural Information Processing Systems (NeurIPS) 2019*, 32, 2019.

Turnbull, B. W. The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3):290–295, 1976.

Wang, H., Talluri, K., and Li, X. On Dynamic Pricing with Covariates. *arXiv preprint arXiv:2112.13254*, 2021a.

Wang, Y., Chen, B., and Simchi-Levi, D. Multimodal Dynamic Pricing. *Management Science*, 67(10):6136–6152, 2021b.

Wang, Y., Chen, X., Chang, X., and Ge, D. Uncertainty Quantification for Demand Prediction in Contextual Dynamic Pricing. *Production and Operations Management*, 30: 1703–1717, 2021c.

Wei, M. M. and Zhang, F. Recent Research Developments of Strategic Consumer Behavior in Operations Management. *Computers & Operations Research*, 93:166–176, 2018.

Xu, J. and Wang, Y.-X. Logarithmic Regret in Feature-based Dynamic Pricing. *Advances in Neural Information Processing Systems (NeurIPS) 2021*, 34, 2021.

Xu, J. and Wang, Y.-X. Towards Agnostic Feature-based Dynamic Pricing : Linear Policies vs Linear Valuation with Unknown Noise. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR 151:9643-9662, 2022.

A. Toy Example

For readers' information, we illustrate a toy example of a customer valuation equipped with the PH model.

Let v_t be the customer valuation on a taxi trip and there is no covariate. Put $p = 3$ dollars, $\Delta p = 0.1$, and $\lambda(p) = 2$. This means that an increase of a taxi trip price from 3 to 3.1 dollars may lead to a loss of approximately $20 \cdot 0.01 = \frac{2}{10}$ of customers who would take a taxi if its price were equal to or less than 3 dollars.

Now, suppose that the customer valuation follows the PH model with $d = 1$ and $\beta = -1$. Let x_t be an indicator of whether t-th customer is experiencing a peak time (e.g. $x_t = 1$ for the rush hours and Friday night and $x_t = 0$ for a normal time). Then by (5), $\lambda(p|x_t = 1)/\lambda(p|x_t = 0) = \exp(-1) \approx 0.37$ for all p. So if $\Delta p = 0.01$ and $\lambda(p|x_t = 0) = 2$, then $\lambda(p|x_t = 1)\Delta p = 2 \times 0.37 \times 0.1 = 0.07$, which means that an increase of taxi trip price from 3 to 3.1 dollars may lead to 7% of customer churn-out in the peak time. As a result, the distribution of v_t on the peak times ($x_t = 1$) will be more concentrated on larger values than that on the normal times ($x_t = 0$).

B. Proofs

B.1. Proof of Lemma 6.2

First, we cite the original statement of Huang (1996). We recall the definition $\|f(\cdot)\|_{L_2(Q)} := \{\int f(u)^2 dQ(u)\}^{\frac{1}{2}}$.

Lemma B.1 (Theorem 3.3 of Huang 1996)**.** *Suppose that $\{(x_t, p_t, y_t)\}_{t=1}^n$ is an i.i.d random sample with $x_t \in \mathbb{R}^d$, $p_t \in [p_{\min}, p_{\max}]$ and $y_t|(p_t, x_t) \sim \text{Bernoulli}(S_0(p_t)^{\exp(x_t^T \beta)})$, where $S_0(\cdot) = 1 - F_0(\cdot)$ for a CDF F_0. Here, $\theta = (\beta, S_0(\cdot))$ is the unknown estimand. If Assumptions 1-3 hold and the joint distribution of (x_t, p_t) does not depend on true θ, then the nonparametric maximum likelihood estimator $\widehat{\theta} = (\widehat{\beta}, \widehat{S}_0(\cdot))$ satisfies*

$$\|\widehat{\beta} - \beta\|_2 + \|\widehat{S}_0(\cdot) - S_0(\cdot)\|_{L_2(Q)} = O_P(n^{-\frac{1}{3}}),$$

where Q is the marginal distribution of p_t.

The dependence on d is hidden in the statement, which we unveil as follows. The original proof uses the empirical process theory, where a key step is obtaining the covering number of the model family. Huang (1996) derived the covering number as $C(1/\epsilon^d)(e^{1/\epsilon})$ for any small $\epsilon > 0$, where $C > 0$ is constant (Lemma 3.1 in the paper). Then this number was further bounded by $C' e^{1/\epsilon}$ for use in subsequent steps. This implies that the log-covering number is approximately $1/\epsilon$; by Lemma A.1 therein, integrating (the square root of) the log-covering number leads to a convergence rate of an empirical process. Note that C' depends on d. We can reveal the dependence on d by bounding $(1/\epsilon^d)(e^{1/\epsilon})$ by $e^{(d+1)/\epsilon}$, i.e., the log-covering number of the family is approximately d/ϵ. Applying the new bound to the remaining steps results in the order of $dn^{-\frac{1}{3}}$. That is, it holds that

$$\|\widehat{\beta} - \beta\|_2 + \|\widehat{S}_0(\cdot) - S_0(\cdot)\|_{L_2(Q)} = O_P(dn^{-\frac{1}{3}}). \tag{15}$$

Coming back to our epoch-based design, we remark that $\widehat{\theta}_k$ uses the data in epoch $k-1$, of which sample size is $|\mathcal{E}_k|/2$. This first implies $n = |\mathcal{E}_k|/2$, so the order $dn^{-\frac{1}{3}}$ in the right-hand side of (15) becomes $d|\mathcal{E}_k|^{-\frac{1}{3}}$. Another implication is that Q in Lemma B.1 in our context is the distribution of p_t in epoch $k-1$, namely Q_{k-1}. That is,

$$\|\widehat{S}_0^k(\cdot) - S_0(\cdot)\|_{L_2(Q_{k-1})} = O_P(d|\mathcal{E}_k|^{-\frac{1}{3}}).$$

By the design, the distribution of p_t at epoch k (Q_k) is continuously supported on $[p_{\min}, p_{\max}]$ for any k. And, by assumption 3d, the distribution of p_t^* on the epoch k (denoted by Q_k^*) is continuously supported on a subset of $(p_{\min}, p_{\max})$. Therefore, we have both Q_k and Q_k^* are absolutely continuous with respect to Q_{k-1}, and the Radon–Nikodym derivatives $|dQ_k/dQ_{k-1}|$ and $|dQ_k^*/dQ_{k-1}|$ are bounded by a constant C. Thus, by the Radon–Nikodym theorem,

$$\|\widehat{S}_0(\cdot) - S_0(\cdot)\|_{L_2(Q_k)} = \int (\widehat{S}_0 - S_0)\frac{dQ_k}{dQ_{k-1}} dQ_{k-1} \le C \int (\widehat{S}_0 - S_0) dQ_{k-1} = O_P(d|\mathcal{E}_k|^{-\frac{1}{3}}),$$

and similarly $\|\widehat{S}_0(\cdot) - S_0(\cdot)\|_{L_2(Q_k^*)} = O_P(d|\mathcal{E}_k|^{-\frac{1}{3}})$. This completes the proof. □

B.2. Proof of Lemma 6.3

Let $t \in \mathcal{E}_k$. For simplicity, we write as $n = |\mathcal{E}_k|$, $\widehat{\beta} = \widehat{\beta}^k$ and $\widehat{S}_0(\cdot) = \widehat{S}_0^k(\cdot)$, if there is no confusion. By algebra, for any $u \in [p_{\min}, p_{\max}]$, $R_{t,k}(u) \le p_{\max} \left|\widehat{S}(u|x_t) - S(u|x_t)\right|$ and

$$\begin{aligned} \left|\widehat{S}(u|x_t) - S(u|x_t)\right| &= \left|\widehat{S}_0(u)^{\exp(x_t^T\widehat{\beta})} - S_0(u)^{\exp(x_t^T\beta)}\right| \\ &\le \left|\widehat{S}_0(u)^{\exp(x_t^T\widehat{\beta})} - S_0(u)^{\exp(x_t^T\widehat{\beta})}\right| + \left|S_0(u)^{\exp(x_t^T\widehat{\beta})} - S_0(u)^{\exp(x_t^T\beta)}\right|. \end{aligned} \tag{16}$$

For the first term of (16), the Mean Value Theorem on a map $t \mapsto t^c$ ($c > 0$ a constant) yields $\left|\widehat{S}_0(u)^{\exp(x_t^T\widehat{\beta})} - S_0(u)^{\exp(x_t^T\widehat{\beta})}\right| = \left|\widehat{S}_0(u) - S_0(u)\right| \cdot \exp(x_t^T\widehat{\beta})\bar{S}(u)^{\exp(x_t^T\widehat{\beta})}$ for some $\bar{S}_0(u)$ between $\widehat{S}_0(u)$ and $S_0(u)$. From the convergence of $\widehat{\beta}$ (Lemma 6.2) and the boundedness of x_t and β (Assumptions 1-2), and $0 < S_0, \widehat{S}_0 < 1$ guarantees $|\exp(x_t^T\widehat{\beta})\bar{S}(u)^{\exp(x_t^T\widehat{\beta})}| < c_1$ for any u for some constant c_1. i.e, the first term of (16) is bounded by $c_1|\widehat{S}_0(u) - S_0(u)|$.

For the second term of (16), the Mean Value theorem on a map $t \mapsto (c')^{\exp(c''t)}$ ($c', c'' > 0$ a constant) yields $\left|S_0(u)^{\exp(x_t^T\widehat{\beta})} - S_0(u)^{\exp(x_t^T\beta)}\right| = |x_t^T(\widehat{\beta} - \beta)| \cdot S_0(u)^{\exp(x_t^T\bar{\beta})}|\log S_0(u)|\exp(x_t^T\bar{\beta})x_t$, for some $\bar{\beta}$ between $\widehat{\beta}$ and β. By the Cauchy-Schwartz inequality $|x_t^T(\widehat{\beta} - \beta)| \le \|x_t\|_2\|\widehat{\beta} - \beta\|_2$, we can make a similar argument with the above paragraph to derive $\left|S_0(u)^{\exp(x_t^T\widehat{\beta})} - S_0(u)^{\exp(x_t^T\beta)}\right| = c_2\|\widehat{\beta} - \beta\|_2$ for a constant c_2.

Letting $C_1 = p_{\max}c_1$ and $C_2 = p_{\max}c_2$, we obtain the desired result. □

B.3. Proof of Lemma 6.4

We want to show (a) $\sum_{t\in\mathcal{E}_k} |\widehat{S}_0^k(p_t) - S_0(p_t)| = O_P(d|\mathcal{E}_k|^{\frac{2}{3}})$; and (b) $\sum_{t\in\mathcal{E}_k} |\widehat{S}_0^k(p_t^*) - S_0(p_t^*)| = O_P(d|\mathcal{E}_k|^{\frac{2}{3}})$. We prove (a) only; then the proof for (b) goes parallel.

For simplicity, let $W_{k,t} = \widehat{S}_0^k(p_t) - S_0(p_t)$, $n = |\mathcal{E}_k|$ and relabel the indices in $\mathcal{E}_k$ as $1, \dots, n$. Now, a decomposition leads to

$$\begin{aligned} \frac{1}{n}\sum_{t=1}^n |W_{k,t}| &= \frac{1}{n}\sum_{t=1}^n \left(|W_{k,t}| - \int |\widehat{S}_0^k(u) - S_0(u)|dQ_k(u) + \int |\widehat{S}_0^k(u) - S_0(u)|dQ_k(u)\right) \\ &= \frac{1}{n}\sum_{t=1}^n \left(|W_{k,t}| - \int |\widehat{S}_0^k(u) - S_0(u)|dQ_k(u)\right) + \int |\widehat{S}_0^k(u) - S_0(u)|dQ_k(u). \end{aligned}$$

For the first term, note that $\{p_t\}_{t=1}^n$ is an i.i.d. sample of Q_k and $\widehat{S}_0^k$ is deterministic given epochs up to $k-1$. And, since $|\widehat{S}_0^k(u) - S_0(u)| \le 2$ for any real number u, any moment of the $W_{k,t}$ is finite. Thus, we can apply the central limit theorem and obtain that the first term is of order $n^{-\frac{1}{2}}$. For the second term, Jensen's inequality leads to $\int |\widehat{S}_0^k(u) - S_0(u)|dQ_k(u) \le \left[\int \{\widehat{S}_0^k(u) - S_0(u)\}^2 dQ_k(u)\right]^{\frac{1}{2}}$, where the right-hand side is of order $dn^{-\frac{1}{3}}$ by Lemma 6.2. Since $dn^{-\frac{1}{3}}$ is a dominating order, we can conclude that $\frac{1}{n}\sum_{t=1}^n |W_{k,t}| = O_P(dn^{-\frac{1}{3}})$, which completes the proof.

□

B.4. Proof of Theorem 6.1

We prove the case without ϵ-greedy heuristic ($\alpha_k = 0$) first and generalize to the case $\alpha_k \ge 0$ in the remark after the proof.

Before proceeding, we note that, with loss of generality, we may assume the last epoch is complete (i.e., $T = \tau_1(2^K - 1)$ for some integer $K \ge 1$). If not (i.e., $\tau_1(2^{K-1} - 1) < T < \tau_1(2^K - 1)$), the regret associated with the incomplete last epoch will be no greater than if it were completed. Thus, the number of epochs K and T satisfies $T = \tau_1(2^K - 1)$, equivalently $K = \log_2(T/\tau_1 + 1)$.

Let $\delta > 0$ be small. From Lemma 6.2, 6.3 and 6.4, there exist $K_0 > 0$ and $C_0, C_1, C_2 > 0$ such that

$$\max\{\|\widehat{\beta}^k - \beta\|_2, \|\widehat{S}_0^k(\cdot) - S_0(\cdot)\|_{L_2(Q_k)}, \|\widehat{S}_0^k(\cdot) - S_0(\cdot)\|_{L_2(Q_k^*)}, J_{2k}, J_{3k}\} \le C_0|\mathcal{E}_k|^{-\frac{1}{3}}d \tag{17}$$

and

$$|R_{t,k}(u)| \le C_1|\widehat{S}_0^k(u) - S_0(u)| + C_2\|\widehat{\beta}^k - \beta\|_2$$

with probability at least $1-\delta$, for any $k > K_0$. In this high-probability event, (11) and (12) leads to

$$regret(k\text{-th epoch}) := \sum_{t\in\mathcal{E}_k} regret(t) \le C_3|\mathcal{E}_k|^{\frac{2}{3}}d, \tag{18}$$

where $C_3 = (2C_1 + 2C_2)C_0^2$. Note that C_3 depends on δ. Thus, we can summarize as follows: For given $\delta > 0$, there exists K_0 and C_3 such that

$$\inf_{k:k>K_0} \mathbb{P}\left[regret(k\text{-th epoch}) \le C_3 d|\mathcal{E}_k|^{\frac{2}{3}}\right] \ge 1-\delta.$$

Then, by union bound arguments, for given small $\delta > 0$ and sufficient large K,

$$\mathbb{P}\left[\bigcap_{k=K_0+1}^{K}\left\{regret(k\text{-th epoch}) \le C_3 d|\mathcal{E}_k|^{\frac{2}{3}}\right\}\right] \ge 1-(K-K_0)\delta > 1-K\delta.$$

Since δ can be arbitrarily small, we can redefine $\frac{\delta}{K}$ as δ. Then,

$$\mathbb{P}\left[\bigcap_{k=K_0+1}^{K}\left\{regret(k\text{-th epoch}) \le C_3 d|\mathcal{E}_k|^{\frac{2}{3}}\right\}\right] \ge 1-\delta \tag{19}$$

for small $\delta > 0$. Let $\mathcal{A}$ be the event in the probability notation.

Now we complete the proof. Suppose that the event $\mathcal{A}$ is true. We decompose by

$$R(T) = \sum_{k=1}^{K_0}\sum_{t\in\mathcal{E}_k} regret(t) + \sum_{k=K_0+1}^{K}\sum_{t\in\mathcal{E}_k} regret(t) =: (I) + (II).$$

For (I), note that $pS_0(p)^{\exp(x^T\beta)}$ is upper bounded by $C_4 := \max\{pS_0(p)^{\exp(v)} : p \in [p_{\min}, p_{\max}], v \in [-\mathcal{B}\mathcal{X}, \mathcal{B}\mathcal{X}]\}$. So

$$(I) \le \sum_{k=1}^{K_0}\sum_{t\in\mathcal{E}_k} C_4 = \tau_1(2^{K_0}-1)C_4,$$

where the right-hand side is finite and does not depend on T. For (II), since $K = \log_2\left(\frac{T}{\tau_1}+1\right)$ and $|\mathcal{E}_k| = \tau_1 2^{k-1}$, under the event $\mathcal{A}$,

$$(II) = \sum_{k=K_0+1}^{\log_2(T/\tau_1+1)} regret(k\text{-th episode}) \le \sum_{k=K_0+1}^{\log_2(T/\tau_1+1)} C_4 d|\mathcal{E}_k|^{\frac{2}{3}} \le C_4 d \sum_{k=1}^{\log_2(T/\tau_1+1)} |\mathcal{E}_k|^{\frac{2}{3}} \le C_5 d T^{\frac{2}{3}}.$$

Therefore, given small $\delta > 0$, we have

$$R(T) \le \tau_1(2^{K_0}-1)C_4 + C_5 d T^{\frac{2}{3}}$$

with probability at least $1-\delta$. As T is sufficiently large, this completes the proof. □

Remark B.2. Now, suppose that there is a ϵ-greedy heuristic with the forced sampling frequency α_k. Then, the total regret at epoch k is decomposed into the regret from rounds with the forced sampling, say r_{1k}, and that from rounds with the myopic policy, say (r_{2k}). Note that $r_{2k} = O_P(2^{\frac{2k}{3}}d)$ from (18). And, since each round of the forced sampling leads to a constant-bounded regret,

$$r_{1k} = O(\alpha_k|\mathcal{E}_k|) = O(\alpha_k|\mathcal{E}_k|).$$

Choosing $\alpha_k = \min\{\gamma 2^{-\frac{k-1}{3}}, 1\}$, we obtain $r_{1k} = O(2^{\frac{2k}{3}})$. Therefore, the regret at the k-th epoch is still $O_P(2^{\frac{2k}{3}}d)$ and we obtain the same result with the case $\alpha_k = 0$.

C. Complete pseudocode of the proposed algorithm (CoxCP) with ϵ-greedy heuristic

Algorithm 2 describes a complete pseudocode of the proposed algorithm (CoxCP) with ϵ-greedy heuristic.

Algorithm 2 The Cox Contextual Pricing (CoxCP) algorithm with ϵ-greedy heuristic

Input: The length of the first epoch, τ_1; the minimum and maximum of price search range, $p_{\min}$ and $p_{\max}$; and the forced sampling frequency $\alpha_k = \min\{\gamma \cdot 2^{-(k-1)/3}, 1\}$
For $t = 1, \ldots, \tau_1$, observe x_t, randomly choose p_t from a distribution supported on $[p_{\min}, p_{\max}]$, and get reward y_t;
for epoch $k = 2, 3, \ldots$ **do**
  Estimate $\theta = (\beta, S_0(\cdot))$ as

$$\widehat{\theta}^k \leftarrow \operatorname*{argmax}_{\theta} L_{k-1}(\theta),$$

  where $L_{k-1}(\theta)$ is defined in (9);
  Set $\tau_k \leftarrow 2\tau_{k-1}$;
  Set $\mathcal{E}_k \leftarrow \{\sum_{r=1}^{k-1} \tau_r + 1, \ldots, \sum_{r=1}^{k} \tau_r\}$;
  for round $t \in \mathcal{E}_k$ **do**
    Observe x_t;
    Draw a binary number r from $\mathrm{Bernoulli}(\alpha_k)$;
    if $r = 0$ **then**

$$p_t \in \operatorname*{argmax}_{p \in [p_{\min}, p_{\max}]} \left\{ p \widehat{S}_0^k(p)^{\exp(x_t^T \widehat{\beta}^k)} \right\}; \tag{20}$$

    else
      Randomly choose p_t from a distribution supported on $[p_{\min}, p_{\max}]$;
    end if
    Get reward y_t.
  end for
end for

D. Details of the simulation

D.1. Configuration

Shah et al. (2019) has four algorithm variants and we ran the DEEP-C policy (Section 3.1 therein) because this version is most highlighted in the main body. The exploration of Shah et al. (2019) can be tuned by the γ (in their notation), the variance of their upper confidence-bound algorithm. We considered $\gamma \in \{3^{-2}, 3^{-1}, 0, 3^1, 3^2\}$. We note that the algorithm discretizes the parameter space at the initial steps, and the number of discretized initial bins is $T^{\frac{d+1}{4}}$. Since the discretization for $T = 30000$ led to $30000^{\frac{5+1}{4}} = 5{,}196{,}152$ initial bins, which was computationally intractable, we ran their algorithms with $\lceil 3000^{\frac{1}{4}} \rceil^{5+1} = 262{,}144$ initial bins.

The exploration of Luo et al. (2022) can be tuned by the length of the first epoch τ_1 and the length of the exploration phase in each epoch. The candidate of τ_1 was $\{64, 128, 256, 512, 1024\}$ as in our algorithm. The length of exploration phase is set $\lceil C_1 l_k^{\beta} \rceil$ in their paper; we considered $C_1 \in \{2^{-2}, 2^{-1}, 1, 2^1, 2^2\}$.

The exploration of Fan et al. (2022) can be tuned in the same way. The length of first epoch τ_1 were set as $\tau_1 \in \{64, 128, 256, 512, 1024\}$. The length of exploration phase in each epoch is $\lceil (l_k d)^{(2m+1)/(4m-1)} \rceil$ in their notation; we considered $C_1 \in \{2^{-2}, 2^{-1}, 1, 2^1, 2^2\}$ as in the same way with Luo et al. (2022).

Both Luo et al. (2022) and Fan et al. (2022)'s algorithms require in common to specify the support of uniform distribution $U[b, B]$ in the "exploration phase" in each episode. We specified b and B as the minimum and maximum of the true support of F.

Other hyperparameters (if any) were set as their default values in their supplemental codes.

Codes for the simulation are available at `https://github.com/younggeunchoi/CoxContextualPricing`.

D.2. Computation time

In Table 2, we present a comparison of computation times for running one instance for $T = 30{,}000$, averaged over settings and replications, measured in seconds.

Table 2. Comparison on computation times for running one instance for $T = 30{,}000$, averaged over settings and replications. Measured in seconds.

ALGORITHMS	COXCP (PROPOSED)	LUO ET AL. (2022)	FAN ET AL. (2022)	SHAH ET AL. (2019)
TIME IN SEC. (AVG. $\pm$ SD.)	20.3 ± 0.4	5.1 ± 0.7	12841.7 ± 36.3	23227.2 ± 224.5

Ours and Luo et al. (2022)'s algorithm showed efficient computation since the computations of parameters are $O(\log T)$ times of convex optimization problem and their price offerings are elementwise products and sums of two vectors. On the other hand, computational inefficiency of Fan et al. (2022) may be mainly due to the evaluation of kernel smoothing estimator, in which the complexity for each t is proportional to the product of the exploration samples and the number of points that one needs to evaluate. In addition, the computational inefficiency of Shah et al. (2019) may be mainly due to the fact that they conduct a $d + 1$-dimensional grid search at initial steps, as explained in the previous subsection.