Mathematics of Operations Research

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

Distribution-Free Contextual Dynamic Pricing

Yiyun Luo, Will Wei Sun, Yufeng Liu

To cite this article:
Yiyun Luo, Will Wei Sun, Yufeng Liu (2023) Distribution-Free Contextual Dynamic Pricing. Mathematics of Operations Research

Published online in Articles in Advance 11 May 2023

. https://doi.org/10.1287/moor.2023.1369

https://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Articles in Advance, pp. 1–20
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Distribution-Free Contextual Dynamic Pricing

Yiyun Luo,[a] Will Wei Sun,[b] Yufeng Liu[c,*]

[a] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China; [b] Krannert School of Management, Purdue University, West Lafayette, Indiana 47907; [c] Departments of Statistics and Operations Research, Genetics, and Biostatistics, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

*Corresponding author

Contact: yiyun851light@gmail.com (YiL); sun244@purdue.edu, https://orcid.org/0000-0002-8412-6430 (WWS); yfliu@email.unc.edu, https://orcid.org/0000-0002-1686-0545 (YuL)

Received: September 8, 2021
Revised: May 31, 2022
Accepted: February 27, 2023
Published Online in Articles in Advance: May 11, 2023

MSC2020 Subject Classification: Primary: 68T05, 62L05, 90B60; secondary: 68Q32

https://doi.org/10.1287/moor.2023.1369

Abstract. Contextual dynamic pricing aims to set personalized prices based on sequential interactions with customers. At each time period, a customer who is interested in purchasing a product comes to the platform. The customer's valuation for the product is a linear function of contexts, including product and customer features, plus some random market noise. The seller does not observe the customer's true valuation, but instead needs to learn the valuation by leveraging contextual information and historic binary purchase feedback. Existing models typically assume full or partial knowledge of the random noise distribution. In this paper, we consider contextual dynamic pricing with unknown random noise in the linear valuation model. Our distribution-free pricing policy learns both the contextual function and the market noise simultaneously. A key ingredient of our method is a novel perturbed linear bandit framework, in which a modified linear upper confidence bound algorithm is proposed to balance the exploration of market noise and the exploitation of the current knowledge for better pricing. We establish the regret upper bound and a matching lower bound of our policy in the perturbed linear bandit framework and prove a sublinear regret bound in the considered pricing problem. Finally, we demonstrate the superior performance of our policy on simulations and a real-life auto loan data set.

Funding: Y. Liu and W.W. Sun acknowledge support from the National Science Foundation Division of Social and Economic Sciences [Grant NSF-SES 2217440].

Supplemental Material: The supplementary material is available at https://doi.org/10.1287/moor.2023.1369.

Keywords: classification • dynamic pricing • linear bandits • regret analysis

1. Introduction

Contextual dynamic pricing aims to design an online pricing policy adaptive to product features, customer characteristics, and the marketing environment (Huang et al. [30]). It is widely used in industries such as hospitality, tourism, entertainment, retail, electricity, and public transportation (den Boer [21]). A successful dynamic pricing algorithm involves both pricing and learning to maximize revenues. Upon receiving sequential customer responses, the algorithm continuously updates its knowledge of customer purchasing behavior and sets a price accordingly. Such online statistical learning differs from traditional supervised or unsupervised learning in its adaptive and sequential manner.

The key learning objective in dynamic pricing is the willingness to pay (demand) of a customer, that is, the probability of a customer making a buying decision. With full knowledge of the demand, the seller can set optimal prices that yield the maximum expected revenues. However, it is common that the seller knows little about the demand prior to the pricing procedure. Such an unknown demand case is studied extensively in dynamic pricing (Besbes and Zeevi [6], Cesa-Bianchi et al. [12], Chen et al. [15], Cheung et al. [17], den Boer and Keskin [23], Keskin and Zeevi [35]). In this case, one critical task is to balance the trade-off between exploration and exploitation, in which exploration aims for more customer-demand knowledge and exploitation maximizes the revenue based on the current knowledge. Two major influential factors for a customer's willingness to pay are the price offered by the seller as well as the customer's valuation of the product. In this paper, we consider a widely adopted linear valuation model (Golrezaei et al. [28], Javanmard and Nazerzadeh [32]). Given the contextual covariate $\boldsymbol{x}$, for example, product features, customer characteristics, and the marketing environment, the customer's valuation $v(\boldsymbol{x})$ for the product is $v(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\theta}_0 + z$. Here, the first component represents the linear effect of the covariates $\boldsymbol{x}$ with an unknown parameter $\boldsymbol{\theta}_0$ and the second component models a market noise z drawn from an unknown distribution

F. After observing the price p set by the seller, the customer buys the product if $v(\boldsymbol{x})$ exceeds p and otherwise leaves without purchasing.

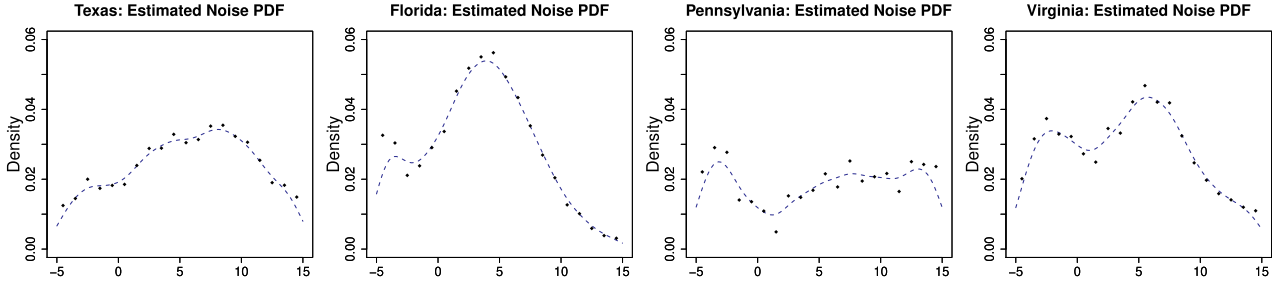
Existing contextual dynamic pricing models assume partial or full knowledge of the market noise distribution F. For example, Javanmard and Nazerzadeh [32] assume a known F for their regularized maximum likelihood pricing (RMLP) method and consider F to belong to a log-concave family for their RMLP-2 policy. Despite knowing that F simplifies the pricing process and improves learning accuracy, it can be restrictive and unrealistic in practice. It is essential to tackle the contextual dynamic pricing problem with an unknown F. Importantly, it may happen in practice that not all relevant contexts can be observed, and such unobserved contexts may lead to a complex noise term. For example, the heterogeneity among customers may lead to a noise that is a mixture of many distributions beyond the log-concave family. In our auto loan data set studied in Section 5, the estimated probability distribution functions (PDFs) of the noise term in four states are clearly not log-concave as shown in Figure 1.

In this paper, we propose a distribution-free pricing (DIP) policy to tackle the contextual dynamic pricing problem with unknown θ_0 and unknown F. DIP employs a doubling trick (Lattimore and Szepesvári [38]) in its framework, which cuts the time horizon into episodes in order to reduce the correlations across data and handle the unknown horizon length. At the beginning of each episode, by formulating the θ_0 estimation into a classification problem in which no prior knowledge of F is required, our DIP policy adopts the logistic regression to estimate θ_0 using data in the previous episode. Given such an estimate, we then translate our single-episode pricing problem into a newly proposed perturbed linear bandit (PLB). PLB can be considered an extension of the classic linear bandit (Abbasi-Yadkori et al. [1], Agrawal and Goyal [2], Chu et al. [19]) and is also of independent interest. Interestingly, the "perturbation level" of the translated PLB can be specified as proportional to the ℓ_1 error of the given θ_0 estimate. A modified linear upper confidence bound (M-LinUCB) algorithm, serving as an essential part of DIP, is proposed for our translated PLB to unify the learning of F and exploitation of the learnt knowledge to set prices.

In addition to the methodological contribution, we also establish regret analysis of our DIP policy. The regret, as the expected loss of revenues with respect to the clairvoyant policy, is widely used to evaluate the performance of a pricing policy. We first prove a T_0-period regret of $\tilde{O}(\sqrt{T_0} + C_p T_0)$ for M-LinUCB on a general PLB with C_p representing the perturbation level. The decomposition of sublinear and linear terms is analogous to the regret in misspecified linear bandits (Foster et al. [27], Lattimore et al. [39], Pacchiano et al. [44]). Importantly, we also show that the linear dependence of T_0 is unavoidable by establishing a matching lower bound for our perturbed linear bandit. We then apply this result to the specific PLB formulation of our single-episode pricing problem to obtain the regret bound for each episode. Finally, we obtain the regret bound for the entire T horizon, which consists of an $\tilde{O}(T^{2/3})$ sublinear term and an extra term related to the estimation error for θ_0. The latter term is dominated by the sublinear term in a broad range of scenarios, which is well-supported by our experiments. In summary, our sublinear $\tilde{O}(T^{2/3})$ regret upper bound implies that the average regret per time period vanishes as the time horizon tends to infinity. Because our problems involve both unknown linear parameter θ_0 and unknown noise distribution F, we conjecture that the obtained $\tilde{O}(T^{2/3})$ rate is close to the optimal rate.

Finally, we demonstrate the superior performance of our policy on extensive simulations and a real-life auto loan data set by comparing our DIP policy to RMLP and RMLP-2 (Javanmard and Nazerzadeh [32]). Because of the restrictive condition on F, RMLP is not satisfactory when a moderate misspecification of F occurs. Despite being more robust than RMLP, RMLP-2 inevitably leads to a linear regret when the noise distribution is beyond log-concave. On the other hand, our DIP policy is robust to unknown complex noise distributions. In a real-life auto loan data set, our DIP policy is shown to largely improve the regret of the benchmark RMLP-2 method in learning customer's purchasing behavior of auto loans. Specifically, DIP has an 80% improvement over RMLP-2 in the

Figure 1. (Color online) Estimated noise PDFs for four states in our auto loan real application.



cumulative regret over the considered time horizon. Such an improvement keeps increasing when the total time horizon increases. See Section 5 for more details.

1.1. Related Work

1.1.1. Noncontextual Dynamic Pricing. For noncontextual dynamic pricing without covariates, Besbes and Zeevi [6, 8], Wang et al. [52], and Chen and Gallego [14] design policies to handle a nonparametric model, whereas Besbes and Zeevi [6], Broder and Rusmevichientong [10], den Boer and Zwart [24], and Keskin and Zeevi [35] consider parametric models. Furthermore, Besbes and Zeevi [7], den Boer [22], and Keskin and Zeevi [36] investigate the time-varying unknown demand setting. In addition, the upper confidence bound (UCB) idea (Abbasi-Yadkori et al. [1], Auer et al. [3]) is used in different noncontextual instances (Kleinberg and Leighton [37], Misra et al. [41], Wang et al. [51]). However, all these approaches do not incorporate the covariates into the pricing policy. Therefore, our model and technical tools are fundamentally different.

1.1.2. Contextual Dynamic Pricing. Dynamic pricing with covariates has garnered significant interest among researchers. As Mueller et al. [42], Javanmard et al. [33], and Chen et al. [16] focus on the multiproduct setting, most of the contextual dynamic pricing literature (Ban and Keskin [4], Bastani et al. [5], Cohen et al. [20], Javanmard [31], Mao et al. [40], Nambiar et al. [43], Qiang and Bayati [47], Wang et al. [53], Xu and Wang [54]) considers a single product at each time. Javanmard and Nazerzadeh [32] and Golrezaei et al. [28, 29] also consider the linear valuation model as we do in this paper. Similar to us, Golrezaei et al. [28] assume both the unknown linear effect and noise distribution and, thus, face the same challenge of error propagation. They adopt a second price auction mechanism with multiple buyers at each time. One main difference lies in the feedback structure. Namely, they assume a full-information setting in which the seller observes all bids and valuations from multiple buyers, whereas we consider a bandit setting in which the seller only observes one single buyer's binary purchasing decision. In Javanmard and Nazerzadeh [32], their proposed RMLP assumes a known market noise distribution, whereas RMLP-2 assumes a known log-concave family of the noise distribution. Hence, their approaches are no longer applicable when the noise distribution is unknown or not log-concave. In addition, by assuming the noise distribution to be in a known ambiguity set, Golrezaei et al. [29] also establish a $\tilde{O}(T^{2/3})$ regret with respect to a robust benchmark defined upon the ambiguity set. In the general unknown noise case, the ambiguity set could be extremely large, and hence, the robust benchmark could be far from the true optimal policy. In contrast, our DIP policy is adaptive to the general unknown noise case, and our regret bound is established by comparing it to the true optimal policy. On the other hand, Shah et al. [49] and Chen and Gallego [13] share similar nonparametric ingredients in the unknown demand function as ours. Specifically, Chen and Gallego [13] consider a general Lipschitz demand and propose a pricing policy based on adaptive binning of the covariate space (Perchet and Rigollet [45]) with a regret of $\tilde{O}(T^{(2+d_0)/(4+d_0)})$, where d_0 is the dimension of covariates. Thus, when $d_0 \geq 3$, our DIP policy enjoys better performance as we leverage the parametric structure in our dynamic pricing model. Shah et al. [49] adopt a log-linear valuation model to handle the unknown nonparametric noise in their semiparametric model. Their method heavily relies on the special structure of the log-linear valuation model, whose optimal price has desirable separable effects of the unknown linear structure and unknown noise distribution. Hence their approach is not applicable to our pricing model in which these two unknown parts tangle with each other. Therefore, techniques used in Shah et al. [49] and Chen and Gallego [13] for handling nonparametric components in the demand function are very different from the newly proposed PLB framework of our DIP policy.

1.1.3. Bandit Algorithms. Our pricing policy is also related to bandit algorithms (Bubeck and Cesa-Bianchi [11], Foster and Rakhlin [26], Lattimore and Szepesvári [38]) which address the balance between exploration and exploitation. In particular, our perturbed linear bandit is related to misspecified linear bandits (Foster et al. [27], Lattimore et al. [39], Pacchiano et al. [44]) and nonstationary linear bandits (Cheung et al. [18], Russac et al. [48], Zhao et al. [56]). An interesting finding is that, by leveraging the special structure of the perturbed linear bandit formulation of our dynamic pricing problem, we achieve a better and more precise regret bound for our proposed policy compared with direct application of much more complex existing algorithms for misspecified or nonstationary linear bandits. See Section 3.1 for more discussions.

1.2. Notation and Paper Organization

We adopt the following notations throughout the article. Let $[T] = \{1, \ldots, T\}$. For a vector $\boldsymbol{\beta} \in \mathbb{R}^d$, let $\|\boldsymbol{\beta}\|_\infty = \max_j |\boldsymbol{\beta}_j|$ and $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^d |\boldsymbol{\beta}_j|$ denote its max norm and ℓ_1 norm, respectively. For two sequences a_n, b_n, we say $a_n = O(b_n)$ if $a_n \leq Cb_n$ for some positive constant C, $a_n = \tilde{O}(b_n)$ if $a_n = O(b_n)$ that ignores logarithmic terms, and $a_n = \Omega(b_n)$ if $a_n \geq Cb_n$ for some positive constant C.

The rest of the paper is organized as follows. In Section 2, we introduce the methodology of our proposed DIP policy along with the perturbed linear bandit formulation of the pricing problem. In Section 3, we develop regret bounds for a general perturbed linear bandit problem and employ it to establish the regret bound of our DIP policy. In Section 4, we demonstrate the superior performance of DIP on various synthetic data sets, and in Section 5, we apply DIP to a real-life auto loan data set. We conclude our work along with some future directions in Section 6. Technical proofs and additional numerical results are collected in the Supplementary Material as an E-companion.

2. Methodology

In this section, we discuss the contextual dynamic pricing problem setting and then introduce our DIP policy, which involves a general perturbed linear bandit formulation.

2.1. Problem Setting

In contextual dynamic pricing, a potential customer who is interested in purchasing a product arrives at the platform at each period $t \in [T] = \{1, \dots, T\}$, and the seller observes a covariate $\boldsymbol{x}_t \in \mathcal{X} \subseteq \mathbb{R}^{d_0}$ representing the product features and customer characteristics. Similar to Javanmard and Nazerzadeh [32], Golrezaei et al. [28], Shah et al. [49], and Chen and Gallego [13], we assume $\|\boldsymbol{x}_t\|_\infty \le 1,\ \forall \boldsymbol{x}_t \in \mathcal{X}$. Given $\boldsymbol{x}_t$, the customer's valuation of the product $v_t = v(\boldsymbol{x}_t) = \boldsymbol{x}_t^\top \boldsymbol{\theta}_0 + z_t$ is a sum of a linear function of $\boldsymbol{x}_t$ and a market noise z_t. We assume $\{z_t\}_{t\in[T]}$ are drawn independent and identically distributed (i.i.d.) from an unknown distribution with cumulative distribution function (CDF) F. If the customer's valuation v_t is higher than the price p_t set by the seller, the sale happens, and the seller collects a revenue of p_t. Otherwise, the customer leaves, and the seller receives no revenue. Let $y_t = 1_{\{v_t \ge p_t\}}$ denote whether the customer buys the product. By the aforementioned sales mechanism, it follows that

$$y_t = \begin{cases} 1 & \text{if } v_t \ge p_t, \text{ with probability } 1 - F(p_t - \boldsymbol{x}_t^\top \boldsymbol{\theta}_0); \\ 0 & \text{if } v_t < p_t, \text{ with probability } F(p_t - \boldsymbol{x}_t^\top \boldsymbol{\theta}_0), \end{cases}$$

and the reward $Z_t = p_t y_t = p_t 1_{\{v_t \ge p_t\}}$. Then, the triplet $(\boldsymbol{x}_t, p_t, y_t)$ records the information of the pricing procedure at time t.

Given this customer choice model and the covariate $\boldsymbol{x}$, the expected reward of setting price p is $p(1 - F(p - \boldsymbol{x}^\top \boldsymbol{\theta}_0))$. We define the optimal price $p^*(\boldsymbol{x})$ as that maximizing $p(1 - F(p - \boldsymbol{x}^\top \boldsymbol{\theta}_0))$, which is an implicit function of the covariate and dependent on both the unknown $\boldsymbol{\theta}_0$ and F. By dynamically setting prices and observing binary feedback, we collect instant revenues and, meanwhile, gather more information to estimate $\boldsymbol{\theta}_0, F$ and $p^*(\boldsymbol{x})$. An important feature of this process is the trade-off between exploration and exploitation in which we well-balance between exploiting the current knowledge for larger immediate revenues and exploring more information for better future revenues.

We next introduce the notion of regret for evaluating a pricing policy. Denote

$$p_t^* = p^*(\boldsymbol{x}_t) = \arg\max_{p>0} p(1 - F(p - \boldsymbol{x}_t^\top \boldsymbol{\theta}_0))$$

as the optimal price at time t. Then, the regret r_t at time t is defined as the loss of reward by setting the price p_t compared with the optimal price p_t^*, that is,

$$r_t = p_t^*(1 - F(p_t^* - \boldsymbol{x}_t^\top \boldsymbol{\theta}_0)) - p_t(1 - F(p_t - \boldsymbol{x}_t^\top \boldsymbol{\theta}_0)). \tag{1}$$

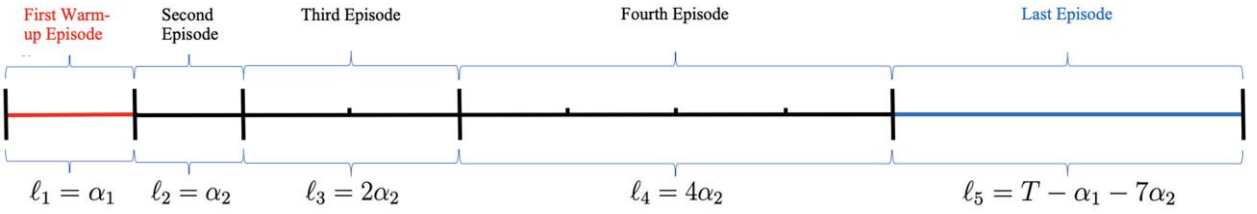
The T-period cumulative regret across the horizon is defined as $R_T = \sum_{t=1}^T r_t$. We obtain the expected cumulative regret $\mathbb{E}(R_T)$ by taking the expectation with respect to the randomness of data and the potential randomness of the pricing policy. The goal of our contextual dynamic pricing is to decide the price p_t for covariate $\boldsymbol{x}_t$ at time t by utilizing all historic data $\{(\boldsymbol{x}_s, p_s, y_s), s = 1, \dots, t-1\}$ in order to minimize the expected cumulative regret.

2.2. DIP Algorithm

Our proposed DIP policy enjoys a simple framework as an outer algorithm nested with inner algorithms A and B. Inner algorithm A is designed for estimating $\boldsymbol{\theta}_0$, and inner algorithm B is the essential part that fully exploits the perturbed linear bandit formulation of our single-episode pricing problem and implements the UCB idea to resolve the trade-off between exploration and exploitation.

2.2.1. Outer Algorithm. In online learning, the total time horizon T is typically unknown. To address this problem, we adopt a doubling trick widely used in online learning and bandit algorithms (Lattimore and Szepesvári [38]) to cut the horizon into episodes. After the first warm-up episode and starting from the second episode, we set the length of the next episode as double the current one until the horizon ends. The number of episodes $n = n(T, \alpha_1, \alpha_2)$

Figure 2. (Color online) An illustration of cutting total time horizon utilizing the doubling trick.



and their lengths denoted as $\{\ell_k = \ell_k(T, \alpha_1, \alpha_2)\}_{k\in[n]}$ are functions of the total horizon length T and the first two episodes' lengths α_1, α_2. Figure 2 demonstrates the case when the total time horizon is cut into five episodes via the doubling trick.

We present the outline of our DIP policy as the generic outer algorithm in Algorithm 1. In the first warm-up episode, DIP performs random exploration to set random prices at each time period. Then DIP alternates between inner algorithm A to obtain an estimate of $\boldsymbol{\theta}_0$ and inner algorithm B to set prices. Specifically, inner algorithm A uses all data from episode $k-1$ to obtain an estimate $\hat{\boldsymbol{\theta}}_{k-1}$ of $\boldsymbol{\theta}_0$; then, inner algorithm B takes $\hat{\boldsymbol{\theta}}_{k-1}$ as an input to sequentially set prices for all time periods in episode k, which then forms all triplets of covariates, prices, and customer responses in episode k for future $\boldsymbol{\theta}_0$ estimation by inner algorithm A. Another advantage of the horizon-cutting strategy is the reduction of correlation across the pricing procedure.

Algorithm 1 (Generic Outer Algorithm)

1: **Input: (arrives over time)** covariates $\{\boldsymbol{x}_t\}_{t\in[T]}$
2: Denote the episodes yielded by the doubling trick as $\mathcal{E}_1, \ldots, \mathcal{E}_n$.
3: **For** $t \in \mathcal{E}_1$, **do**
4: Set a price p_t randomly from $(0, p_{\max})$ and receive a binary response y_t.
5: **For** episode $k = 2, 3, \ldots, n$, **do**
6: With input data $\{(\boldsymbol{x}_t, p_t, y_t)\}_{t\in\mathcal{E}_{k-1}}$, apply inner algorithm A on this data set to update an estimate $\hat{\boldsymbol{\theta}}_k$ of $\boldsymbol{\theta}_0$;
7: With input $\hat{\boldsymbol{\theta}}_{k-1}$ as the estimate of $\boldsymbol{\theta}_0$, apply inner algorithm B on $\mathcal{E}_k$ to sequentially set a price p_t and receive a binary response y_t for all $t \in \mathcal{E}_k$.

2.2.2. Inner Algorithm A. We now introduce the inner algorithm A designed for estimating $\boldsymbol{\theta}_0$. It uses all data $(\boldsymbol{x}_t, p_t, y_t)$ from the $(k-1)$th episode to obtain an estimate $\hat{\boldsymbol{\theta}}_{k-1}$ for future pricing in the kth episode. For simplicity, we introduce its generic version with $[T_0] = \{1, \ldots, T_0\}$ representing the $(k-1)$th episode horizon. Because y_t is binary and invoked by $\boldsymbol{x}_t, p_t$ through $\mathbb{P}(y_t = 1) = 1 - F(p_t - \boldsymbol{x}_t^\top \boldsymbol{\theta}_0)$, we obtain

$$\begin{cases} \mathbb{P}(y_t = 1) > \dfrac{1}{2}, \text{ if } F^{-1}\left(\dfrac{1}{2}\right) + \boldsymbol{x}_t^\top \boldsymbol{\theta}_0 - p_t > 0; \\ \mathbb{P}(y_t = 1) = \dfrac{1}{2}, \text{ if } F^{-1}\left(\dfrac{1}{2}\right) + \boldsymbol{x}_t^\top \boldsymbol{\theta}_0 - p_t = 0; \\ \mathbb{P}(y_t = 1) < \dfrac{1}{2}, \text{ if } F^{-1}\left(\dfrac{1}{2}\right) + \boldsymbol{x}_t^\top \boldsymbol{\theta}_0 - p_t < 0. \end{cases}$$

Therefore, we can form a classification problem with responses y_t and covariates $(1, \boldsymbol{x}_t^\top, p_t)^\top$ for $t \in [T_0]$. It admits a Bayes decision boundary $\{u : (F^{-1}(1/2), \boldsymbol{\theta}_0^\top, -1)u = 0\}$, which involves the unknown parameter $\boldsymbol{\theta}_0$. Thus, we can estimate the linear decision boundary and extract an estimate of $\boldsymbol{\theta}_0$ by applying a linear classification method. In this paper, we use logistic regression, which yields an estimate $(\hat{c}, \hat{\boldsymbol{\beta}}^\top, \hat{b})$ of $(F^{-1}(1/2), \boldsymbol{\theta}_0^\top, -1)$ up to a constant factor. Thus, $-\hat{\boldsymbol{\beta}}/\hat{b}$ is a natural estimate of $\boldsymbol{\theta}_0$. Similar to Javanmard and Nazerzadeh [32], we assume $\|\boldsymbol{\theta}_0\|_1$ is upper bounded by a known constant W. By projecting $-\hat{\boldsymbol{\beta}}/\hat{b}$ onto the ℓ_1-ball $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^{d_0} : \|\boldsymbol{\theta}\|_1 \le W\}$, we can obtain our final estimate denoted as $\hat{\boldsymbol{\theta}} = \mathrm{Proj}_\Theta(-\hat{\boldsymbol{\beta}}/\hat{b})$. Such a projection has a closed-form solution as $\mathrm{Proj}_\Theta(-\hat{\boldsymbol{\beta}}/\hat{b}) = \mathcal{T}_{\rho_{\min}}(-\hat{\boldsymbol{\beta}}/\hat{b})$, where $\mathcal{T}_\rho(v) = \mathrm{sgn}(v)(|v| - \rho)_+$ is the soft-thresholding operator and $\rho_{\min} = \min\{\rho : \left\|\mathcal{T}_\rho(-\hat{\boldsymbol{\beta}}/\hat{b})\right\|_1 \le W\}$. Here, the assumption of constant W is purely for theoretical purposes, and our policy is very robust to the value of W in the empirical studies. The generic inner algorithm A is summarized in Algorithm 2.

Algorithm 2 (Generic Inner Algorithm A)

1: **Input:** $\{(\boldsymbol{x}_t, p_t, y_t)\}_{t\in[T_0]}$, W
2: Use logistic regression to obtain the minimizer

$$(\hat{c}, \hat{\boldsymbol{\beta}}^\top, \hat{b}) = \arg\min_{(c,\beta^\top,b)} \sum_{t=1}^{T_0} \log(1+\exp((2y_t-1)(c,\beta^\top,b)(1,\boldsymbol{x}_t^\top,p_t)^\top)).$$

3: Estimate $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}} = \mathrm{Proj}_\Theta(-\hat{\boldsymbol{\beta}}/\hat{b})$, where $\Theta = \{\boldsymbol{\theta}\in\mathbb{R}^{d_0} : \|\boldsymbol{\theta}\|_1 \le W\}$.

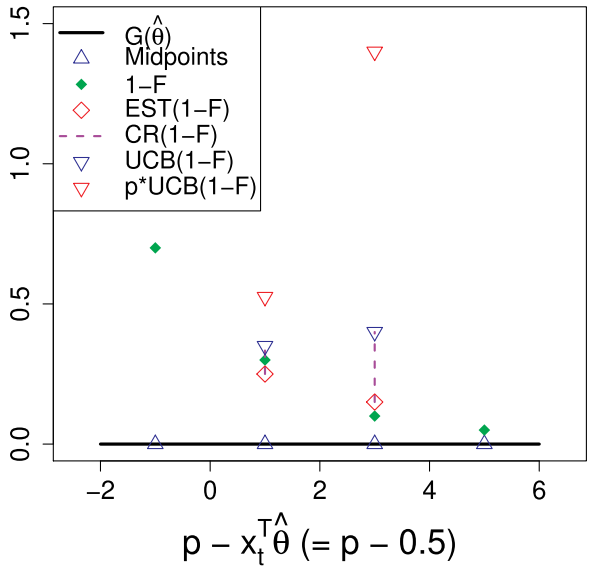
Under the same assumption of a known upper bound W of $\|\boldsymbol{\theta}_0\|_1$, RMLP and RMLP-2 in Javanmard and Nazerzadeh [32] estimate $\boldsymbol{\theta}_0$ via the maximum likelihood type of method by assuming some knowledge on F. In comparison, our approach achieves robust $\boldsymbol{\theta}_0$ estimation without knowledge of a potentially complex-shaped F. It is worth mentioning that the logistic regression used in Algorithm 2 can be replaced by other linear classification methods, for example, large-margin classifiers (Wang et al. [50]). We choose logistic regression for its simplicity and superior numerical performance.

2.2.3. Inner Algorithm B. Next, we introduce the inner algorithm B designed for setting prices. Taking $\hat{\boldsymbol{\theta}}_{k-1}$ obtained by inner algorithm A as an input, it sequentially sets prices for all time periods in episode k. For ease of presentation, we introduce a generic version by using $\hat{\boldsymbol{\theta}}$ to represent $\hat{\boldsymbol{\theta}}_{k-1}$ and T_0 to represent the length of the episode k.

Based on our model in Section 2.1, the knowledge of the expected reward $p(1-F(p-\boldsymbol{x}_t^\top\boldsymbol{\theta}_0))$ plays a critical role in deciding the best price at time t. Given the current estimate $\hat{\boldsymbol{\theta}}$, we need to evaluate $\{p(1-F(p-\boldsymbol{x}_t^\top\hat{\boldsymbol{\theta}}))\}$ over $p\in(0,p_{\max})$. Here, we assume there is a known upper bound $p_{\max}$ of our pricing problem. This assumption is very mild in real applications and is also used in Javanmard and Nazerzadeh [32] and Chen and Gallego [13]. By the condition $\|\boldsymbol{x}_t\|_\infty \le 1$, we have $p-\boldsymbol{x}_t^\top\hat{\boldsymbol{\theta}} \in G(\hat{\boldsymbol{\theta}}) = [-\|\hat{\boldsymbol{\theta}}\|_1, p_{\max}+\|\hat{\boldsymbol{\theta}}\|_1]$. Therefore the evaluation of the expected reward is reduced to evaluate $1-F$ on $G(\hat{\boldsymbol{\theta}})$. When F is Lipschitz continuous and no other global smoothness is assumed, it is sufficient to evaluate $1-F$ on several well-chosen discrete points in $G(\hat{\boldsymbol{\theta}})$ to leverage the finite data for better pricing. In this paper, we utilize the discretization idea (Kleinberg and Leighton [37]) to cut $G(\hat{\boldsymbol{\theta}})$ into d same-length subintervals with the set of their midpoints $\mathcal{M} = \{m_1,\dots,m_d\}$. Here, d is a parameter that possibly depends on the horizon length T_0. When T_0 is large, it is reasonable to set a larger d for a denser discretization and, hence, larger exploration spaces. We leave the detailed discussion on the choice of d to the theoretical analysis of DIP in Section 3. Our aim is then to dynamically set prices and evaluate $1-F$ on $\mathcal{M}$.

Toy Example (Discretization). We introduce a toy example to better illustrate our pricing policy. We couple each part of our pricing strategy with its corresponding realization in this toy example. All quantities that are introduced in our pricing policy for this specific example are displayed in Figure 3. Consider a two-dimensional covariate

Figure 3. An illustration of inner algorithm B via Toy Example.



$\boldsymbol{x}_t = (0.3, 0.2)^\top$ at time t. Assume we have an estimation $\hat{\boldsymbol{\theta}} = (1,1)^\top$ and $p_{\max} = 4$. Then, the interval for discretization is $G(\hat{\boldsymbol{\theta}}) = [-\|\hat{\boldsymbol{\theta}}\|_1, p_{\max} + \|\hat{\boldsymbol{\theta}}\|_1] = [-2, 6]$ represented by the black solid line in Figure 3. If $d = 4$, we discretize $G(\hat{\boldsymbol{\theta}})$ into subintervals $[-2, 0], [0, 2], [2, 4]$, and $[4, 6]$. Their midpoints $m_1 = -1, m_2 = 1, m_3 = 3, m_4 = 5$, represented by blue hollow triangles on the black line in Figure 3, form the set $\mathcal{M} = \{-1, 1, 3, 5\}$. We continue this example later.

To achieve the mutual reinforcement of pricing and evaluation of $1 - F$ on $\mathcal{M}$, we restrict the set price p_t at time t into a carefully constructed candidate set $\mathcal{S}_t = \{m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} | j \in [d], m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} \in (0, p_{\max})\}$. The key feature for any price $p \in \mathcal{S}_t$ is that $p - \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}$ exactly equals a midpoint in $\mathcal{M}$. We now illustrate why pricing in $\mathcal{S}_t$ and evaluation of $1 - F$ on $\mathcal{M}$ can enhance each other. For any price $p = m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} \in \mathcal{S}_t$, we can leverage our current knowledge of $1 - F(m_j)$ to obtain an estimate of its expected reward $p(1 - F(p - \boldsymbol{x}_t^\top \boldsymbol{\theta}_0))$ as $p(1 - F(p - \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}})) = p(1 - F(m_j))$. Thus, a better evaluation of $1 - F$ on $\mathcal{M}$ improves our pricing decision from $\mathcal{S}_t$. On the other hand, when we set one price $p_t = m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}$ from $\mathcal{S}_t$, we observe a binary response $y_t \sim \text{Ber}(1 - F(m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} - \boldsymbol{x}_t^\top \boldsymbol{\theta}_0)) \approx \text{Ber}(1 - F(m_j))$, which then improves our knowledge of $1 - F(m_j)$. Upon this observation, we say that we pull arm j at time t if we set $p_t = m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}$. Then, pulling arm j yields more knowledge for $1 - F$ on m_j. Thus, we define the available arm set at time t as $\mathcal{B}_t = \{j \in [d] : \exists p \in \mathcal{S}_t \text{ such that } p = m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}\}$, which varies over time as $\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}$ changes over time.

Toy Example (Continued, Construct Candidate Sets). We construct the candidate sets $\mathcal{S}_t$ based on the discretized set $\mathcal{M} = \{m_1 = -1, m_2 = 1, m_3 = 3, m_4 = 5\}$. As $\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} = 0.5$, we obtain $\mathcal{S}_t = \{m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} | m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} \in (0, p_{\max})\} = \{m_2 + 0.5, m_3 + 0.5\} = \{1.5, 3.5\}$ because $m_1 + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} = -0.5$ and $m_4 + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} = 5.5$ are out of the range $(0, p_{\max})$. In this case, the arm set at time t is $\mathcal{B}_t = \{j \in [d] : \exists p \in \mathcal{S}_t \text{ such that } p = m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}\} = \{2, 3\}$.

Restricted on $\mathcal{S}_t$, there is a clear trade-off between exploration and exploitation for our pricing problem. A pure exploration tends to pull less-pulled arms in $\mathcal{B}_t$ and may set many suboptimal prices, whereas a pure exploitation may continuously pull suboptimal arms because of a lack of knowledge of other arms. To balance between exploration and exploitation, we utilize the principle of optimism in the face of uncertainty (Lattimore and Szepesvári [38]) to construct an upper confidence bound, which calls for both an estimation $\text{EST}_t(1 - F(m_j))$ for $1 - F(m_j)$ and a confidence radius (CR) $\text{CR}_t(1 - F(m_j))$ of this estimation at the beginning of time t. We can accomplish this goal using all the past data yielded by pulling arm j. We leave the specific forms of $\text{EST}_t(1 - F(m_j))$ and $\text{CR}_t(1 - F(m_j))$ to the next section as they emerge naturally from the perturbed linear bandit formulation of our single-episode pricing problem. Then, we select $p_t = m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} \in \mathcal{S}_t$ with the largest optimism estimation $p_t \text{UCB}_t(1 - F(m_j))$, where $\text{UCB}_t(1 - F(m_j)) = \text{EST}_t(1 - F(m_j)) + \text{CR}_t(1 - F(m_j))$ is an optimism estimation of $1 - F(m_j)$. This optimism estimation addresses the exploration–exploitation trade-off because a large UCB can result in either exploring a less-pulled arm with a large CR or exploiting an optimal arm with a large mean estimation.

Toy Example (Continued, Set Prices). As the available arm set is $\mathcal{B}_t = \{2, 3\}$ at time t, we only require knowledge of $1 - F(m_2)$ and $1 - F(m_3)$ to compare between two candidate prices $m_2 + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}$ and $m_3 + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}$. To emphasize this, in Figure 3, we only show $\{\text{EST}_t(1 - F(m_j))\}_{j=2,3}$ (red hollow diamonds) and $\{\text{CR}_t(1 - F(m_j))\}_{j=2,3}$ (lengths of purple dashed line) at two midpoints $m_2 = 1$ and $m_3 = 3$. Summing them up leads to the optimism estimations $\{\text{UCB}_t(1 - F(m_j))\}_{j=2,3}$ represented by blue hollow inverted triangles. Multiplying them by their corresponding prices $m_2 + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} = 1.5$ and $m_3 + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} = 3.5$, we obtain their optimism expected reward estimations represented by red hollow inverted triangles, which are used to form our pricing decisions. Based on the illustration in Figure 3, we set the price $p_t = 3.5$, that is, $m_3 + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}$, because $1.5\text{UCB}_t(1 - F(m_2)) < 3.5\text{UCB}_t(1 - F(m_3))$.

We summarize the generic inner algorithm B for one episode in Algorithm 3.

Algorithm 3 (Generic Inner Algorithm B)

Input: (arrives over time) covariates $\{\boldsymbol{x}_t\}_{t \in [T_0]}$, $\hat{\boldsymbol{\theta}}$, discretization number d, and other inputs required to construct the specific forms of $\{\text{UCB}_t(1 - F(m_j))\}_{j \in [d]}$.
Cut the interval $G(\hat{\boldsymbol{\theta}}) = [-\|\hat{\boldsymbol{\theta}}\|_1, p_{\max} + \|\hat{\boldsymbol{\theta}}\|_1]$ into d same-length intervals and denote their midpoints as $m_1, \ldots, m_d$.
For time $t = 1, \ldots, T_0$, **do**
  Construct the candidate price set $\mathcal{S}_t = \{m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} | j \in [d], m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} \in (0, p_{\max})\}$;
  Determine the arm set $\mathcal{B}_t = \{j \in [d] : \exists p \in \mathcal{S}_t \text{ such that } p = m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}\}$;
  Calculate $\text{UCB}_t(1 - F(m_j))$ for $j \in \mathcal{B}_t$ in (2);
  Calculate $j_t \in \arg\max_{j \in \mathcal{B}_t} (m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}) \text{UCB}_t(1 - F(m_j))$;
  Set a price $p_t = m_{j_t} + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}$ and receive a binary response y_t.

2.2.4. Perturbed Linear Bandit. In this section, we first introduce a PLB framework and then show that our single-episode pricing problem can be formulated as a PLB. Furthermore, the proposed M-LinUCB for PLB is shown to be equivalent to the inner algorithm B with a specific UCB construction.

We say that the reward Z_t, the parameter $\boldsymbol{\xi}_t$, and the action set $\mathcal{A}_t$ form a perturbed linear bandit with a perturbation constant C_p if $Z_t = \langle \boldsymbol{\xi}_t, \boldsymbol{A}_t \rangle + \eta_t$ with any selected action $\boldsymbol{A}_t \in \mathcal{A}_t$ and $\|\boldsymbol{\xi}_s - \boldsymbol{\xi}_t\|_\infty \le C_p$ for any s, t. Here, η_t is a sub-Gaussian conditional on the filtration $\mathcal{F}_{t-1} = \sigma(\boldsymbol{\xi}_1, \boldsymbol{A}_1, Z_1, \ldots, \boldsymbol{\xi}_t, \boldsymbol{A}_t)$. Note that the condition on the linear parameters $\boldsymbol{\xi}_t$'s implies the existence of a $\boldsymbol{\xi}^*$ such that $\|\boldsymbol{\xi}_t - \boldsymbol{\xi}^*\|_\infty \le \frac{C_p}{2}$ for any t. Thus, the linear parameter $\boldsymbol{\xi}_t$ regulating the reward structure at time t can be viewed as a perturbation from a "central" parameter $\boldsymbol{\xi}^*$. Note that the linear bandit (Abbasi-Yadkori et al. [1], Agrawal and Goyal [2], Chu et al. [19]) is a special zero-perturbation PLB with $\boldsymbol{\xi}_t = \boldsymbol{\xi}^*$ for any t.

Now, we introduce the perturbed linear bandit formulation of our single-episode pricing problem with time horizon $[T_0]$. We first specify the linear parameter $\boldsymbol{\xi}_t = (1 - F(m_1 + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} - \boldsymbol{x}_t^\top \boldsymbol{\theta}_0), \ldots, 1 - F(m_d + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} - \boldsymbol{x}_t^\top \boldsymbol{\theta}_0))^\top \in \mathbb{R}^d$, which turns out to regulate the reward at time t as shown in Lemma 1. Note that, for any price $m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} \in \mathcal{S}_t$, the jth element of $\boldsymbol{\xi}_t$ is exactly the purchasing probability of the customer faced with this price. Further define $\boldsymbol{\xi}^* = (1 - F(m_1), \ldots, 1 - F(m_d))^\top$ as the central parameter. Then, by Lemma 1, $\boldsymbol{\xi}_t$'s can be viewed as perturbations from $\boldsymbol{\xi}^*$. It is interesting to see that the perturbations indeed originate from the difference between the estimate $\hat{\boldsymbol{\theta}}$ and the true $\boldsymbol{\theta}_0$ and may change with covariates $\boldsymbol{x}_t$'s.

To transform price setting into action selection, we define a mapping from any price $p = m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} \in \mathcal{S}_t$ to a vector $Q_t(p) \in \mathbb{R}^d$ with $Q_t(p)_j = m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}$ and $Q_t(p)_i = 0$, $\forall i \ne j$. Namely, Q_t maps a price $p = m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} \in \mathcal{S}_t$ to a vector with a single nonzero jth element p. Further define a vector set $\mathcal{A}_t = \{Q_t(p) : p \in \mathcal{S}_t\}$. Then, Q_t is a one-to-one mapping from S_t to $\mathcal{A}_t$, and Q_t^{-1} is well-defined. To proceed, we define the price–action coupling by $\boldsymbol{A}_t = Q_t(p_t)$. Then, setting any price $p_t \in \mathcal{S}_t$ means selecting an action $\boldsymbol{A}_t = Q_t(p_t) \in \mathcal{A}_t$ and vice versa. With all these preparations, the following Lemma 1 rigorously forms our single-episode pricing problem into a perturbed linear bandit given Assumption 1 that assumes a Lipschitz F.

Assumption 1. *F is Lipschitz with the Lipschitz constant L.*

Lemma 1. *Under Assumption 1, $\|\boldsymbol{\xi}_t - \boldsymbol{\xi}^*\|_\infty \le L\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1$, $\forall t \in [T_0]$. Moreover, under the price–action coupling $\boldsymbol{A}_t = Q_t(p_t)$, the reward $Z_t = p_t 1_{\{v_t \ge p_t\}}$, the parameter $\boldsymbol{\xi}_t$, and the action set $\mathcal{A}_t$ form a perturbed linear bandit with a perturbation constant $2L\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1$.*

Lemma 1 implies that the perturbation is proportional to the ℓ_1 estimation error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1$. If the estimate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$, then $\boldsymbol{\xi}_t = \boldsymbol{\xi}^*$ with zero perturbation, and the PLB reduces to a classic linear bandit. On the other hand, a worse $\hat{\boldsymbol{\theta}}$ implies a larger perturbation, thus incurring more difficulty in solving the PLB and potentially leading to a larger regret.

According to Lemma 1, $Z_t = \boldsymbol{A}_t^\top \boldsymbol{\xi}_t + \eta_t$ with $\|\boldsymbol{\xi}_t - \boldsymbol{\xi}^*\|_\infty \le L\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1$, and hence, $\boldsymbol{\xi}^*$ can be estimated from historical data. Similar to that in linear bandit (Lattimore and Szepesvári [38]), we employ the ridge estimator $\hat{\boldsymbol{\xi}}_{t-1} = V_{t-1}(\lambda)^{-1} \sum_{s=1}^{t-1} \boldsymbol{A}_s Z_s$, where $V_{t-1}(\lambda) = \lambda I + \sum_{s=1}^{t-1} \boldsymbol{A}_s \boldsymbol{A}_s^\top$ with the tuning parameter $\lambda > 0$. Note that, in Algorithm 3, we use j_t to denote the arm pulled at time t. Let $\mathcal{U}_{t-1,j} = \{s : 1 \le s \le t-1, j_s = j\}$. Because $\boldsymbol{A}_s$s have a single nonzero element and $V_{t-1}(\lambda)$ is a diagonal matrix, we obtain the explicit form for the jth element of $\hat{\boldsymbol{\xi}}_{t-1}$ as $\hat{\boldsymbol{\xi}}_{t-1,j} = \left(\sum_{s \in \mathcal{U}_{t-1,j}} p_s^2 y_s\right) \Big/ \left(\lambda + \sum_{s \in \mathcal{U}_{t-1,j}} p_s^2\right)$, which serves as the estimate $\mathrm{EST}_t(1 - F(m_j))$ for $1 - F(m_j) = \boldsymbol{\xi}_j^*$.

In order to construct a UCB using the principle of optimism in the face of uncertainty, we then compute a confidence radius $\mathrm{CR}_t(1 - F(m_j))$ of the preceding estimate $\mathrm{EST}_t(1 - F(m_j)) = \hat{\boldsymbol{\xi}}_{t-1,j}$. The common confidence set $\mathcal{C}_t(\beta_t) = \{\boldsymbol{\xi} \in \mathbb{R}^d : \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{t-1}\|^2_{V_{t-1}(\lambda)} \le \beta_t\}$ yields a marginal confidence radius for each $\hat{\boldsymbol{\xi}}_{t-1,j}$. Because of the simple form of $V_{t-1}(\lambda)$, we obtain an explicit form $\mathrm{CR}_t(1 - F(m_j)) = \sqrt{\beta_t / (\lambda + \sum_{s \in \mathcal{U}_{t-1,j}} p_s^2)}$. Then, we obtain the UCB as required in inner algorithm B,

$$\mathrm{UCB}_t(1 - F(m_j)) = \frac{\sum_{s \in \mathcal{U}_{t-1,j}} p_s^2 y_s}{\lambda + \sum_{s \in \mathcal{U}_{t-1,j}} p_s^2} + \sqrt{\frac{\beta_t}{\lambda + \sum_{s \in \mathcal{U}_{t-1,j}} p_s^2}}. \tag{2}$$

Motivated by the linear bandit (Lattimore and Szepesvári [38]), we specify the parameter $\beta_t = \beta_t^* = p_{\max}^2 \left(1 \vee \left(1/p_{\max}\sqrt{\lambda d} + \sqrt{2\log(1/\delta) + d\log\left((d\lambda + (t-1)p_{\max}^2)/d\lambda\right)}\right)^2\right)$. Here, $1 - \delta$ is the confidence level, and $\delta = 1/T_0$ is a typical choice (Lattimore and Szepesvári [38]) with known T_0. Thus, we use $\delta = 1/(2^{k-2}\ell_2)$ for the application of

inner algorithm B to the kth episode with a projected length of $2^{k-2}\ell_2$. Now, we are ready to present the full version of our DIP policy as Algorithm 4. In summary, DIP well-organizes two subalgorithms across episodes, one applying classification for linear parameter estimation and the other adapting the UCB idea for online pricing.

Algorithm 4 (DIP for Contextual Dynamic Pricing)

Input: (at time 0) $\alpha_1, \alpha_2, p_{\max}, C, \lambda, W$
Input: (arrives over time) covariates $\{\boldsymbol{x}_t\}_{t\in[T]}$
For time $t = 1, \ldots, \ell_1(= \alpha_1)$, **do**
  Set a price p_t randomly from $(0, p_{\max})$ and receive a binary response y_t.
For episodes $k = 2, 3, \ldots, n(= n(T, \alpha_1, \alpha_2))$, **do**
  Apply inner algorithm A with the input data $\{(\boldsymbol{x}_t, p_t, y_t)\}_{\sum_{i=1}^{k-2}\ell_i+1\le t\le\sum_{i=1}^{k-1}\ell_i}$ and W to obtain the estimate $\hat{\boldsymbol{\theta}}_{k-1}$;
  Apply inner algorithm B on the coming sequential covariates $\{\boldsymbol{x}_t\}_{\sum_{i=1}^{k-1}\ell_i+1\le t\le\sum_{i=1}^{k}\ell_i}$ with the estimate $\hat{\boldsymbol{\theta}}_{k-1}$, discretization number $d_k = C\lceil(2^{k-2}\ell_2)^{(1/6)}\rceil$, and the UCB construction in (2) with $\beta_t = \beta_t^*$ and $\delta = 1/(2^{k-2}\ell_2)$.

Remark 1. In this remark, we provide the computational complexity of Algorithm 4. In each episode k, the inner algorithm A consists of a logistic regression and a projection with the complexity of $O(\ell_k d_0)$ and $O(d_0)$, respectively. Thus, it contributes a complexity of $O(d_0 T + d_0 \log T) = O(d_0 T)$ in the total horizon. The inner algorithm B in episode k first conducts a discretization with the complexity of $O(d_k) = O(\ell_k^{1/6})$. At its tth iteration, by saving related quantities, the update of constructed upper confidence bounds for $\{1 - F(m_j)\}_{j\in[d_k]}$ takes only $O(1)$ time complexity. The calculation of the estimated linear valuation component $\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_k$ has a complexity of $O(d_0)$. The calculation of optimism expected revenues of the candidate prices and the selection of an optimal candidate price take another $O(d_k)$ time complexity. Thus, the overall complexity of inner algorithm B in episode k is $O(d_0\ell_k + d_k\ell_k) = O(d_0\ell_k + \ell_k^{(7/6)})$. Thus, inner algorithm B contributes a total complexity of $O(d_0 T + T^{(7/6)})$ to the entire horizon. Hence, the computational complexity of the whole DIP policy is $O(d_0 T + T^{(7/6)})$.

Finally, we mention that the proposed perturbed linear bandit framework can be used beyond the contextual dynamic pricing problem. This motivates us to introduce a general algorithm called M-LinUCB in Algorithm 5 for the perturbed linear bandit framework $Z_t = \langle \boldsymbol{\xi}_t, \boldsymbol{A}_t\rangle + \eta_t$ when any potential action has only one nonzero element. For any vector $\boldsymbol{v}$ with a single nonzero element, denote $\delta(\boldsymbol{v})$ as the index of this nonzero element. For instance, $\delta((0,1,0)^\top) = 2$. Further define $\tilde{\mathcal{B}}_t = \{\delta(\boldsymbol{a}) : \boldsymbol{a} \in \mathcal{A}_t\}$ as the nonzero index set of all potential actions at time t and $\tilde{\mathcal{B}}_t' = \{\delta(\boldsymbol{A}_s) : s \in [t-1]\}$ as the nonzero index set of all past selected actions. Then, bridged by the PLB formulation of our single-episode pricing problem, there exists a close connection between M-LinUCB and inner algorithm B formalized in Lemma 2.

Algorithm 5 (M-LinUCB for Perturbed Linear Bandit)

Input: (arrives over time) action sets $\mathcal{A}_t$, λ, $\{\beta_t\}_{t\in[T_0]}$
For $t = 1, \ldots, T_0$, **do**
  Determine $\tilde{\mathcal{B}}_t = \{\delta(\boldsymbol{a}) : \boldsymbol{a} \in \mathcal{A}_t\}$ and $\tilde{\mathcal{B}}_t' = \{\delta(\boldsymbol{A}_s) : s \in [t-1]\}$.
  If $\tilde{\mathcal{B}}_t \nsubseteq \tilde{\mathcal{B}}_t'$, **do**
    Choose an arbitrary $\boldsymbol{A}_t \in \mathcal{A}_t$ such that $\delta(\boldsymbol{A}_t) \notin \tilde{\mathcal{B}}_t'$.
  If $\tilde{\mathcal{B}}_t \subseteq \tilde{\mathcal{B}}_t'$, **do**
    For $\boldsymbol{a} \in \mathcal{A}_t$, **do**
      Calculate $\mathrm{LinUCB}_t(\boldsymbol{a}) = \max_{\boldsymbol{\xi}\in\mathcal{C}_t(\beta_t)}\langle \boldsymbol{\xi}, \boldsymbol{a}\rangle$, where $\mathcal{C}_t(\beta_t) = \{\boldsymbol{\xi} \in \mathbb{R}^d : \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{t-1}\|^2_{V_{t-1}(\lambda)} \le \beta_t\}$ and $\hat{\boldsymbol{\xi}}_{t-1} = V_{t-1}(\lambda)^{-1}$ $\sum_{s=1}^{t-1}\boldsymbol{A}_s Z_s$, $V_{t-1}(\lambda) = \lambda I + \sum_{s=1}^{t-1}\boldsymbol{A}_s\boldsymbol{A}_s^\top$.
  Choose $\boldsymbol{A}_t \in \arg\max_{\boldsymbol{a}\in\mathcal{A}_t}\mathrm{LinUCB}_t(\boldsymbol{a})$.
  Receive a reward Z_t.

Lemma 2. *Applying Algorithm 5 to the PLB formulation of our single-episode pricing problem with* $\beta_t = \beta_t^* = p_{\max}^2 \left(1 \vee \left((1/p_{\max})\sqrt{\lambda d} + \sqrt{2\log(1/\delta) + d\log\left((d\lambda + (t-1)p_{\max}^2)/d\lambda\right)}\right)^2\right)$ *yields Algorithm 3 using the UCB construction (2) with* $\beta_t = \beta_t^*$.

Therefore, inner algorithm B (Algorithm 3) can be viewed as the "projection" of M-LinUCB onto our single-episode pricing problem. In the remaining part of this paper, without further specifications, we refer to Algorithms 3 and 5 as the ones mentioned in Lemma 2.

3. Theory

In this section, we establish the regret bound for the proposed DIP policy. As DIP divides the total time horizon into episodes, we conduct the regret analysis on a single episode and then merge them together. For the single-episode pricing problem, our discretization procedure leads to a natural decomposition of the regret into discrete and continuous parts. One key technical contribution is the proof of the discrete-part regret, which is shown via the equivalent regret of M-LinUCB for the corresponding PLB formulation.

In our single-episode regret analysis, we denote the total horizon as $[T_0]$ and use $\hat{\boldsymbol{\theta}}$ as the input for Algorithm 3. In Algorithm 3, we restrict the price in a discrete candidate set $\mathcal{S}_t$, thus yielding a "discrete" best price $\tilde{p}_t^*$ in $\mathcal{S}_t$, that is, $\tilde{p}_t^* \in \arg\max_{p\in\mathcal{S}_t} p(1-F(p-\boldsymbol{x}_t^\top\boldsymbol{\theta}_0))$. Thus, the regret r_t in (1) can be rewritten as

$$\underbrace{\tilde{p}_t^*(1-F(\tilde{p}_t^*-\boldsymbol{x}_t^\top\boldsymbol{\theta}_0))-p_t(1-F(p_t-\boldsymbol{x}_t^\top\boldsymbol{\theta}_0))}_{r_{t,1}}+\underbrace{p_t^*(1-F(p_t^*-\boldsymbol{x}_t^\top\boldsymbol{\theta}_0))-\tilde{p}_t^*(1-F(\tilde{p}_t^*-\boldsymbol{x}_t^\top\boldsymbol{\theta}_0))}_{r_{t,2}}.$$

The first part $r_{t,1}$ is the reward loss with respect to the discrete best price $\tilde{p}_t^*$. The second part $r_{t,2}$ is the regret of setting $\tilde{p}_t^*$. Denote their sums as $R_{T_0,1}=\sum_{t=1}^{T_0} r_{t,1}$ and $R_{T_0,2}=\sum_{t=1}^{T_0} r_{t,2}$, which are the discrete- and continuous-part regrets, respectively. Then, bounding the cumulative regret $R_{T_0}=R_{T_0,1}+R_{T_0,2}$ reduces to bounding $R_{T_0,1}$ and $R_{T_0,2}$ separately. As discussed before, the discrete-part regret is shown to be the same as the regret under the equivalent PLB formulation and then investigated by utilizing newly developed regret bounds for the PLB setting. For the continuous-part regret, we adopt the following second order smoothness assumption on the general expected revenue function defined as $f_q(p)=p(1-F(p-q))$. Note that the single-step continuous-part regret $r_{t,2}$ can then be rewritten as $f_{\boldsymbol{x}_t^\top\boldsymbol{\theta}_0}(p_t^*)-f_{\boldsymbol{x}_t^\top\boldsymbol{\theta}_0}(\tilde{p}_t^*)$.

Assumption 2. *There exists a constant C such that, for any $q=\boldsymbol{x}^\top\boldsymbol{\theta}_0$ and $\boldsymbol{x}\in\mathcal{X}$, we have $f_q(p^*(\boldsymbol{x}))-f_q(p)\le C(p^*(\boldsymbol{x})-p)^2$, $\forall p\in[0,p_{\max}]$.*

Assumption 2 requires that the reward difference between the overall best price and any other price can be bounded by a constant multiplying their quadratic difference. Given the global continuity of F, Assumption 2 indicates a uniform control of $f_{\boldsymbol{x}^\top\boldsymbol{\theta}_0}(p)$ over the local neighborhoods of the maximizers $p^*(\boldsymbol{x})$. In Proposition 1, by applying Taylor's theorem with the Lagrange remainder, we provide a sufficient condition for Assumption 2. Nevertheless, Assumption 2 does not require any global smoothness of F. The derived regret bound still holds for locally erratic Fs as long as Assumption 2 is satisfied.

Proposition 1. *Assumption 2 holds if $F''(\cdot)$ is bounded on $[-\|\boldsymbol{\theta}_0\|_1, p_{\max}+\|\boldsymbol{\theta}_0\|_1]$.*

Now, we present our main result in the following Theorem 1. It provides a regret upper bound over the entire horizon.

Theorem 1. *Under Assumptions 1 and 2, the DIP policy yields the expected regret*

$$\mathbb{E}(R_T)=\tilde{O}(T^{2/3})+4p_{\max}L\sum_{k=2}^{n}2^{k-2}\ell_2\mathbb{E}\|\hat{\boldsymbol{\theta}}_{k-1}-\boldsymbol{\theta}_0\|_1.$$

Theorem 1 demonstrates how the estimation errors $\|\hat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_0\|_1$ affect the regret upper bound for DIP policy. If the estimates $\{\hat{\boldsymbol{\theta}}_k\}_{k\in[n-1]}$ are perfectly accurate, the second term vanishes, and the overall regret is $\tilde{O}(T^{2/3})$. In general, if $\mathbb{E}\|\hat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_0\|_1=O(\ell_k^{-\alpha})$ for some $0<\alpha\le 1/2$, we can conclude that $\sum_{k=2}^{n}2^{k-2}\ell_2\mathbb{E}\|\hat{\boldsymbol{\theta}}_{k-1}-\boldsymbol{\theta}_0\|_1=O(T^{1-\alpha})$ by the doubling construction. Then, the overall regret is $\tilde{O}(T^{(2/3)\vee(1-\alpha)})$. Because we use the adaptive pricing data in the previous episode to estimate $\boldsymbol{\theta}_0$, it is challenging to derive the exact rate of convergence for the estimation. In spite of this theoretical difficulty, we conduct a simulation study in Section 4.1 to numerically demonstrate that the convergence rate of $\|\hat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_0\|_1$ is between $1/3$ and $1/2$, and hence, the $\tilde{O}(T^{2/3})$ overall regret bound can be practically achieved.

Remark 2. At first glance, the obtained regret upper bound is worse than the typical $\Omega(T^{1/2})$ lower bound in linear bandit (Lattimore and Szepesvári [38]) and dynamic pricing with the known noise distribution (Javanmard and Nazerzadeh [32]). However, we point out that our problem involves both unknown linear parameter $\boldsymbol{\theta}_0$ and unknown noise distribution F. We conjecture that our obtained regret upper bound is close to the lower bound in our setting. To see it, Chen and Gallego [13] consider a nonparametric pricing problem and prove an $\Omega(T^{(d_0+2)/(d_0+4)})$ lower bound under some additional smoothness assumptions, where d_0 is the dimension of the nonparametric component. With a one-dimensional nonparametric component F in our pricing problem, their results suggest an

$\Omega(T^{3/5})$ lower bound, which is higher than the typical $\Omega(T^{1/2})$ rate. However, their constructed instances do not fit into our considered pricing problem with an additional linear structure $\boldsymbol{x}^\top \boldsymbol{\theta}_0$. The additional unknown $\boldsymbol{\theta}_0$ makes the lower bound derivation harder, and we leave it for future work.

Remark 3. After our initial submission, there are two recent papers (Fan et al. [25], Xu and Wang [55]) considering a similar dynamic pricing problem with the unknown noise distribution. In Fan et al. [25], the authors consider an $m(\geq 2)$ times continuously differentiable F and propose an explore-then-commit type of policy that achieved an $\tilde{O}(T^{(2m+1)/(4m-1)})$ regret upper bound. Both our assumed Lipschitz and second order smoothness assumptions are satisfied under their condition of twice continuously differentiable F ($m = 2$). Thus, even under stronger assumptions, their proved $\tilde{O}(T^{5/7})$ regret for $m = 2$ is still worse than our main $\tilde{O}(T^{2/3})$ regret term. In Xu and Wang [55], the authors consider an adversarial setting and propose a D2-EXP4 policy that achieves a regret of $\tilde{O}(T^{3/4})$. By fully utilizing the smoothness of the noise distribution, our proposed DIP policy achieves an improvement to $\tilde{O}(T^{2/3})$ for our main regret term.

In the next two sections, we first do some preparations by developing the regret bounds for the general perturbed linear bandit. Then, we provide a proof outline for our main Theorem 1 by utilizing the proved PLB results.

3.1. Regret Bounds for Perturbed Linear Bandit

We consider a PLB setting with the reward model $Z_t = \langle \boldsymbol{\xi}_t, \boldsymbol{A}_t \rangle + \eta_t$, which satisfies the following conditions.

Condition 1. *For any $t \in \mathbb{N}^+$ and $\boldsymbol{a} \in \mathcal{A}_t$, $|\langle \boldsymbol{\xi}_t, \boldsymbol{a} \rangle| \leq 1$.*

Condition 2. *For any $t \in \mathbb{N}^+$, $\|\boldsymbol{\xi}_t\|_\infty \leq C_1$.*

Condition 3. *For any $t \in \mathbb{N}^+$ and $\boldsymbol{a} \in \mathcal{A}_t$, $\|\boldsymbol{a}\|_0 = 1$ and $\|\boldsymbol{a}\|_2 \leq a_{\max}$ for a constant $a_{\max}$.*

Condition 4. *For any $t \in \mathbb{N}^+$, η_t is a 1-conditionally sub-Gaussian random variable, that is, $\mathbb{E}(\exp(\alpha \eta_t) | \mathcal{F}_{t-1}) \leq \exp(\alpha^2/2)$ for any $\alpha \in \mathbb{R}$, where $\mathcal{F}_{t-1} = \sigma(\boldsymbol{\xi}_1, \boldsymbol{A}_1, Z_1, \ldots, \boldsymbol{\xi}_t, \boldsymbol{A}_t)$.*

Remark 4. Condition 1 ensures a constant regret upper bound at each time and is commonly adopted in linear bandit (Lattimore and Szepesvári [38]). Condition 2 assumes the bounded infinity norm of $\boldsymbol{\xi}_t$. Condition 3 implies there is only one nonzero element bounded in absolute value for any action. This holds for our PLB formulation becausee any action vector $Q_t(p)$ with $p = m_j + \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} \in \mathcal{S}_t$ has a single nonzero jth element $p \in (0, p_{\max})$. Condition 4 implies that the noise is sub-Gaussian conditional on all the past parameters, actions, and rewards as well as the current parameter and action. The perturbed linear bandit formulation of our single-episode pricing problem satisfies all these conditions.

We develop the following Lemma 3 to establish the regret bound for such a PLB setting.

Lemma 3. *Consider the PLB satisfying Conditions 1–4 with a perturbation C_p. With probability at least $1 - \delta$, Algorithm 5 with $\beta_t = \tilde{\beta}_t = 1 \vee \left(C_1 \sqrt{\lambda d} + \sqrt{2\log(1/\delta) + d\log\left((d\lambda + (t-1)a_{\max}^2)/d\lambda\right)} \right)^2$ has the regret bound*

$$R_{T_0}^{PLB} \leq 2\sqrt{2dT_0\tilde{\beta}_{T_0}\log\left(\frac{d\lambda + T_0 a_{\max}^2}{d\lambda}\right)} + 2\boldsymbol{a}_{\max} C_p T_0 + 2d.$$

Proof Sketch. We construct a new sequence of "shadow" linear parameters $\{\dot{\boldsymbol{\xi}}_t\}_{2 \leq t \leq T_0}$ and control the "pseudo-regret" $\sum_{t=2}^{T_0} \langle \dot{\boldsymbol{\xi}}_t, \dot{\boldsymbol{A}}_t - \boldsymbol{A}_t \rangle$ with the sublinear order $\tilde{O}(\sqrt{T_0})$, where $\dot{\boldsymbol{A}}_t = \arg\max_{\boldsymbol{a} \in \mathcal{A}_t} \langle \dot{\boldsymbol{\xi}}_t, \boldsymbol{a} \rangle$. By proving closeness of $\dot{\boldsymbol{\xi}}_t$ and $\boldsymbol{\xi}_t$ for all t, we can bound the difference between the true regret and pseudo-regret by a linear term proportional to the perturbation C_p. The detailed construction of $\{\dot{\boldsymbol{\xi}}_t\}_{2 \leq t \leq T_0}$ and rigorous proofs are deferred to Supplemental Section A. □

As shown in Lemma 3, the second term in the regret upper bound is proportional to the perturbation C_p. When $C_p = 0$, this linear term vanishes, and the final regret bound matches that of the classic linear bandit. Reversely, the perturbed linear bandit becomes intractable when C_p is too large. Interestingly, by Lemma 1, this perturbation constant in the PLB formulation of our single-episode pricing problem is proportional to $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1$. This matches the intuition that a larger estimation error leads to more revenue loss. Lemma 3 is of independent interest because it provides an informative regret bound for the PLB problem.

Remark 5. Our proposed PLB can be viewed as a misspecified linear bandit (Foster et al. [27], Lattimore et al. [39], Pacchiano et al. [44]) with a misspecification level $\epsilon_* = \boldsymbol{a}_{\max} C_p/2$, where the latter has a general regret of $\tilde{O}(d\sqrt{T_0} + \epsilon_* \sqrt{d} T_0)$. In comparison, by leveraging Condition 3, we prove a regret of $\tilde{O}(d\sqrt{T_0} + \boldsymbol{a}_{\max} C_p T_0) = \tilde{O}(d\sqrt{T_0} + \epsilon_* T_0)$ for the simple M-LinUCB algorithm under our PLB setting. To see it, the key in our proof is the closeness property $\|\dot{\boldsymbol{\xi}}_t - \boldsymbol{\xi}_t\|_\infty \le C_p$ of our constructed shadow parameters $\dot{\boldsymbol{\xi}}_t = V_{t-1}^+ \sum_{s=1}^t \boldsymbol{A}_s \boldsymbol{A}_s^\top \boldsymbol{\xi}_s$. The proof of this property relies on the fact that the Moore–Penrose inverse V_{t-1}^+ of $V_{t-1} = V_{t-1}(0)$ is a diagonal matrix, which is a direct result of the condition $\|\boldsymbol{a}\|_0 = 1$ in Condition 3. Importantly, this $\sqrt{d}$ improvement is critical for us to derive the final regret rate of our pricing problem.

Remark 6. Nonstationary linear bandits (NLBs) (Cheung et al. [18], Russac et al. [48], Zhao et al. [56]) also allow changing linear parameters $\boldsymbol{\xi}_t$ but design policies to adapt to the smooth variations $B_{T_0} = \sum_{t=1}^{T_0-1} \|\boldsymbol{\xi}_t - \boldsymbol{\xi}_{t+1}\|_2$. Our PLB setting fits an NLB with linear variations $B_{T_0} = O(C_p T_0)$. The nonasymptotic results in Cheung et al. [18] and Zhao et al. [56] suggest a regret of $\tilde{\mathcal{O}}(B_{T_0}^{1/3} T_0^{2/3}) = \tilde{\mathcal{O}}(C_p^{1/3} T_0)$, which is only valid for a range of C_p (exclusive of zero and dependent on T_0). In contrast, our proven Lemma 3 provides regret behaviors with a fixed T_0 for $C_p \to 0$, that is, approaching the classic linear bandit result $\tilde{\mathcal{O}}(\sqrt{T_0})$ linearly with C_p, which is essential for further derivations in our pricing problem. Though some intermediate results in Cheung et al. [18] and Zhao et al. [56] also yield regrets for fixed T_0 and $C_p \to 0$, they suggest worse regrets, such as $\tilde{\mathcal{O}}(wC_pT_0 + (T_0/\sqrt{w}))$ (w chosen from $\{1, \dots, T_0\}$) and $\tilde{\mathcal{O}}(C_p T_0^2 + \sqrt{T_0})$ when applied to our PLB setting, which inevitably deteriorates the performance guarantee for our pricing problem.

Next, we prove an $\Omega(C_p T_0)$ regret lower bound for the PLB with a perturbation C_p. This implies that the linear term in the upper bound is inevitable because of the potentially adversarial perturbations. Define $PB(\tilde{\boldsymbol{\xi}}, C_p) = \{\boldsymbol{\xi} \in \mathbb{R}^d : \|\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}\|_\infty \le (C_p/2)\}$ as a parameter set with respect to a central parameter $\tilde{\boldsymbol{\xi}}$ and a perturbation quantification C_p.

Proposition 2. *For any PLB algorithm $\mathcal{A}^*$, any $\tilde{\boldsymbol{\xi}}$ with all positive elements, and $(C_p/2) < \min_{i\in[d]} \tilde{\boldsymbol{\xi}}_i$, there exists a PLB with parameters $(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_t, \dots)$ and action sets $(\mathcal{A}_1, \dots, \mathcal{A}_t, \dots)$ satisfying $\boldsymbol{\xi}_t \in PB(\tilde{\boldsymbol{\xi}}, C_p)$, $\forall t \in \mathbb{N}^+$ and a constant C_0 only dependent on $\tilde{\boldsymbol{\xi}}$ such that*

$$\mathbb{E}(R_{T_0}^{PLB}(\mathcal{A}^*)) \ge C_0 C_p T_0, \ \forall T_0 \in \mathbb{N}^+.$$

3.2. Proof Outline for Theorem 1

To prove Theorem 1, we first prove regret upper bounds for each episode and then merge them together. In the following Proposition 3, we prove a high-probability regret bound as well as an expected regret bound for our pricing policy in a single episode. Specifically, the expected regret is bounded by a sublinear $\tilde{O}(T^{2/3})$ term and a linear term proportional to the ℓ_1 estimation error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1$. As DIP applies Algorithm 3 to the kth episode with $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{k-1}$, we obtain Theorem 1 by applying Proposition 3 to each episode.

Proposition 3. *Under Assumptions 1 and 2 with probability at least $1 - \delta$, applying Algorithm 3 on the single-episode pricing problem yields the total regret R_{T_0} satisfying*

$$R_{T_0} \le 2\sqrt{2dT_0\beta_{T_0}^* \log\left(\frac{d\lambda + T_0 p_{\max}^2}{d\lambda}\right)} + 4p_{\max} L \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 T_0 + C_0 \frac{T_0}{d^2} + 2dp_{\max}.$$

Moreover, by setting $\delta = (1/T_0), d = C\lceil T_0^{1/6} \rceil$ and taking the expectation, we have $\mathbb{E}(R_{T_0}) = \tilde{O}(T_0^{2/3}) + 4p_{\max} L \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 T_0$.

It remains to prove the regret bound in Proposition 3 for the single-episode pricing problem. We conduct the analysis of both the discrete- and continuous-part regret and then combine them together.

3.2.1. Discrete-Part Regret. By the PLB formulation in Lemma 1 and the one-to-one correspondence between $\mathcal{S}_t$ and $\mathcal{A}_t$, the best action in $\mathcal{A}_t$ is $\boldsymbol{A}_t^* = Q_t(\tilde{p}_t^*)$. Therefore, the selected action $\boldsymbol{A}_t = Q_t(p_t)$ yields the regret $\tilde{p}_t^*(1 - F(\tilde{p}_t^* - \boldsymbol{x}_t^\top \boldsymbol{\theta}_0)) - p_t(1 - F(p_t - \boldsymbol{x}_t^\top \boldsymbol{\theta}_0))$ for the PLB, which matches the discrete-part regret $r_{t,1}$. Moreover, Lemma 2 shows that Algorithm 5 yields Algorithm 3 under the price–action coupling. Therefore, we can investigate the regret of Algorithm 5 on the PLB to quantify the discrete-part regret of Algorithm 3.

We now apply the general regret bound of Lemma 3 to the PLB formulation of our single-episode pricing problem to bound the discrete-part regret. After scaling the rewards, linear parameters and noises by $1/p_{\max}$ as $\tilde{\boldsymbol{\xi}}_t = (1/p_{\max})\boldsymbol{\xi}_t$, $\tilde{Z}_t = (1/p_{\max})Z_t$, $\tilde{\eta}_t = (1/p_{\max})\eta_t$, we obtain the transformed model $\tilde{Z}_t = \langle \tilde{\boldsymbol{\xi}}_t, \boldsymbol{A}_t \rangle + \tilde{\eta}_t$ with the perturbation

constant $\tilde{C}_p = 2L\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1/p_{\max}$, which satisfies Conditions 1–4 with $C_1 = 1/p_{\max}$ and $a_{\max} = p_{\max}$. On the other hand, we can prove that applying Algorithm 5 with $\beta_t = \beta_t^*$ on the original PLB is equivalent to applying it with $\beta_t = \tilde{\beta}_t = (1/p_{\max}^2)\beta_t^*$ on the transformed model with their regrets admitting a scaling relationship. By formalizing this reasoning, we obtain the following Proposition 4.

Proposition 4. *Under Assumption 1 with probability at least* $1-\delta$*, applying Algorithm 3 on the single-episode pricing problem yields a discrete-part regret* $R_{T_0,1}$ *satisfying*

$$R_{T_0,1} \le 2\sqrt{2dT_0\beta_{T_0}^*\log\left(\frac{d\lambda + T_0p_{\max}^2}{d\lambda}\right)} + 4p_{\max}L\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 T_0 + 2dp_{\max}.$$

Proposition 4 provides an upper bound of the discrete-part regret on a single episode. The first term is sublinear as $\tilde{O}(\sqrt{T_0})$, whereas the second term is linear in T_0 and proportional to the estimation error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1$, which invokes the perturbation in our PLB formulation. The third term is dominated by the first two terms as we further specify d to yield a best trade-off between discrete and continuous parts of the regret.

3.2.2. Continuous-Part Regret. We now discuss how to derive a bound for the continuous-part regret under Assumption 2. By our discretization approach, $\{m_i + \boldsymbol{x}_t^\top\hat{\boldsymbol{\theta}}\}_{i\in[d]}$ are a sequence of points that "cover" $[0, p_{\max}]$ with equal adjacent distance $(p_{\max} + 2\|\hat{\boldsymbol{\theta}}\|_1)/d$. Because $\mathcal{S}_t = \{m_j + \boldsymbol{x}_t^\top\hat{\boldsymbol{\theta}} \,|\, j\in[d], m_j + \boldsymbol{x}_t^\top\hat{\boldsymbol{\theta}} \in (0, p_{\max})\}$ and $p_t^* \in (0, p_{\max})$, there must exist a $\dot{p}_t \in \mathcal{S}_t$ close enough with p_t^* such that their expected reward difference is $O(1/d^2)$ according to Assumption 2. Because the discrete best price $\tilde{p}_t^*$ outperforms $\dot{p}_t$, the unit continuous-part regret $r_{t,2}$ of setting $\tilde{p}_t^*$ satisfies $r_{t,2} = O(1/d^2)$. Thus, the continuous-part regret $R_{T_0,2}$ in the entire horizon is of the order $O(T_0/d^2)$.

3.2.3. Combination. We can prove that the right-hand side of the regret result in Proposition 4 has a simpler form of $\tilde{O}(d\sqrt{T_0}) + 4p_{\max}L\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 T_0$. Thus, the overall regret for the single episode is $\tilde{O}\left(d\sqrt{T_0} + (T_0/d^2)\right) + 4p_{\max}L\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 T_0$. By setting the discretization number d in the order of $T_0^{1/6}$, we obtain the single-episode regret bounds in Proposition 3.

4. Simulation Study

We demonstrate the performance of our DIP policy on synthetic data sets and compare it with RMLP and RMLP-2 proposed by Javanmard and Nazerzadeh [32]. The implementation details of DIP, RMLP, and RMLP-2 are provided in Supplemental Section B.

Let $\Phi(\mu, \sigma^2)$ denote the CDF of $N(\mu, \sigma^2)$ distribution. For the first six examples, we consider a scalar covariate $\boldsymbol{x}_t \overset{\text{i.i.d.}}{\sim} \text{Unif}[0,1]$ and set $\boldsymbol{\theta}_0 = 30$. The CDF F of the noise distribution is designed as follows, in which Examples 1 and 5 are motivated from the real application in Section 5. Their probability density functions are shown in Figure 4.

Example 1. The true $F = (1/2)\Phi(-4, 6) + (1/2)\Phi(4, 6)$.

Example 2. The true $F = (1/3)\Phi\left(-6, (\pi^2/3)\right) + (1/3)\Phi\left(-1, (\pi^2/3)\right) + (1/6)\Phi\left(1, (\pi^2/3)\right) + (1/6)\Phi\left(6, (\pi^2/3)\right)$.

Example 3. The true $F = (1/4)\Phi\left(-7, (\pi^2/3)\right) + (1/4)\Phi\left(-3, (\pi^2/3)\right) + (1/4)\Phi\left(3, (\pi^2/3)\right) + (1/4)\Phi\left(7, (\pi^2/3)\right)$.

Example 4. The true $F(\cdot) = \tilde{F}(\cdot + \text{mean}(\tilde{F}))$, where $\tilde{F} = (1/3)\Phi\left(-3, (\pi^2/3)\right) + (2/3)\Phi\left(3, (\pi^2/3)\right)$.

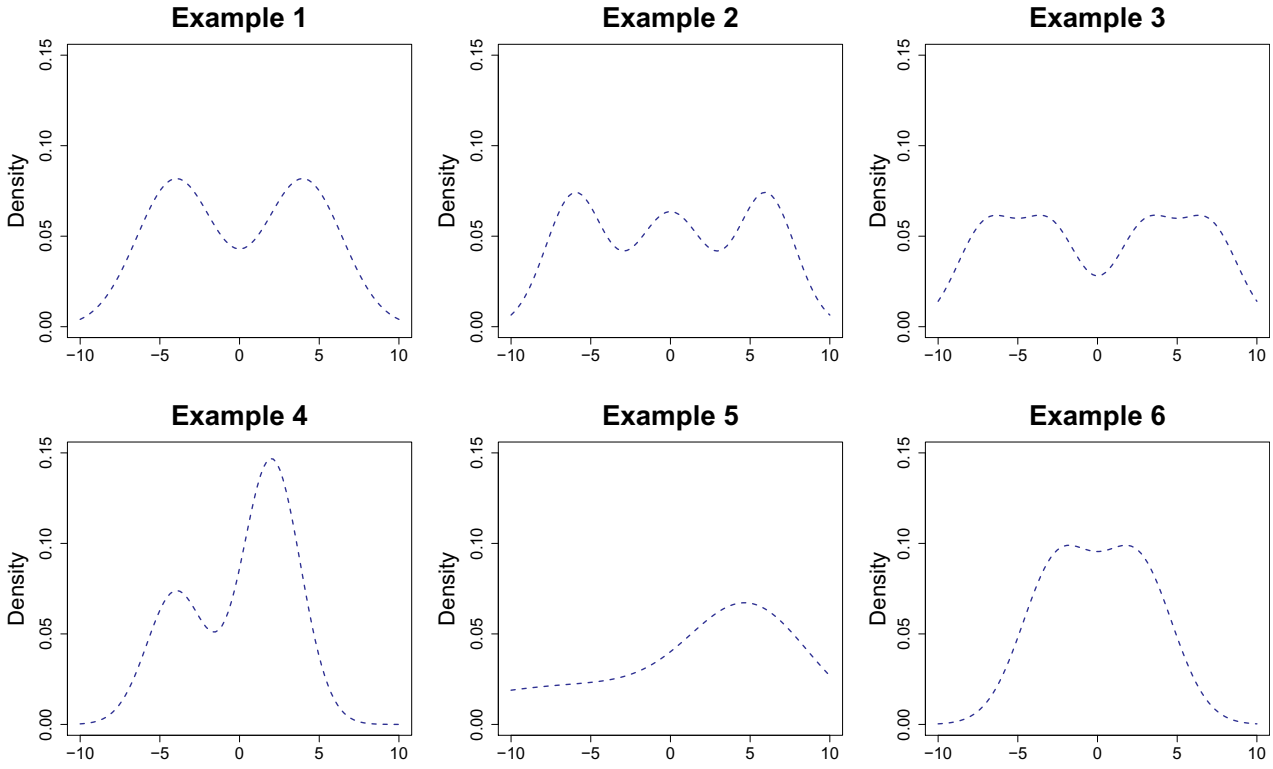
Example 5. The true $F(\cdot) = \tilde{F}(\cdot + \text{mean}(\tilde{F}))$, where $\tilde{F} = (1/2)\Phi\left(-5, (25\pi^2/3)\right) + (1/2)\Phi\left(5, (4\pi^2/3)\right)$.

Example 6. The true $F = (1/2)\Phi(-2.5, 5) + (1/2)\Phi(2.5, 5)$.

As shown in Figure 4, Examples 1–3 have symmetric PDFs with two, three, and four modes, respectively, whereas Examples 4 and 5 have asymmetric PDFs with two modes and one single mode, respectively. Example 6 has two peaks but is close to a single-mode normal distribution.

We compute the mean and confidence interval of cumulative regrets over 100 replications. As shown in Figure 5, DIP outperforms both RMLP and RMLP-2 for Examples 1–5. In Example 6, RMLP and RMLP-2 perform better than DIP as the noise distribution F is close to a normal distribution, which aligns with their model assumption. Because of the misspecification of F, the cumulative regrets for both RMLP and RMLP-2 exhibit clear linear patterns. The performance deterioration of RMLP and RMLP-2 becomes severe in Example 5, in which the PDF is asymmetric and has heavy tails. On the other hand, our DIP policy gradually learns F in the

Figure 4. (Color online) PDFs of the noise distribution in Examples 1–6.



pricing process and achieves sublinear cumulative regrets in all examples. These examples illustrate the severity of noise distribution misspecification of RMLP and RMLP-2 and, hence, the superior performance of the proposed robust pricing policy.

Next, we show that DIP can still outperform RMLP and RMLP-2 even when the noise distribution F is standard Gaussian $\Phi(0,1)$, which satisfies the log-concave condition assumed by RMLP-2. In the following Examples 7–9, we set $F = \Phi(0,1)$ and vary the context dimension d_0, the true $\boldsymbol{\theta}_0$, and the context generation distribution.

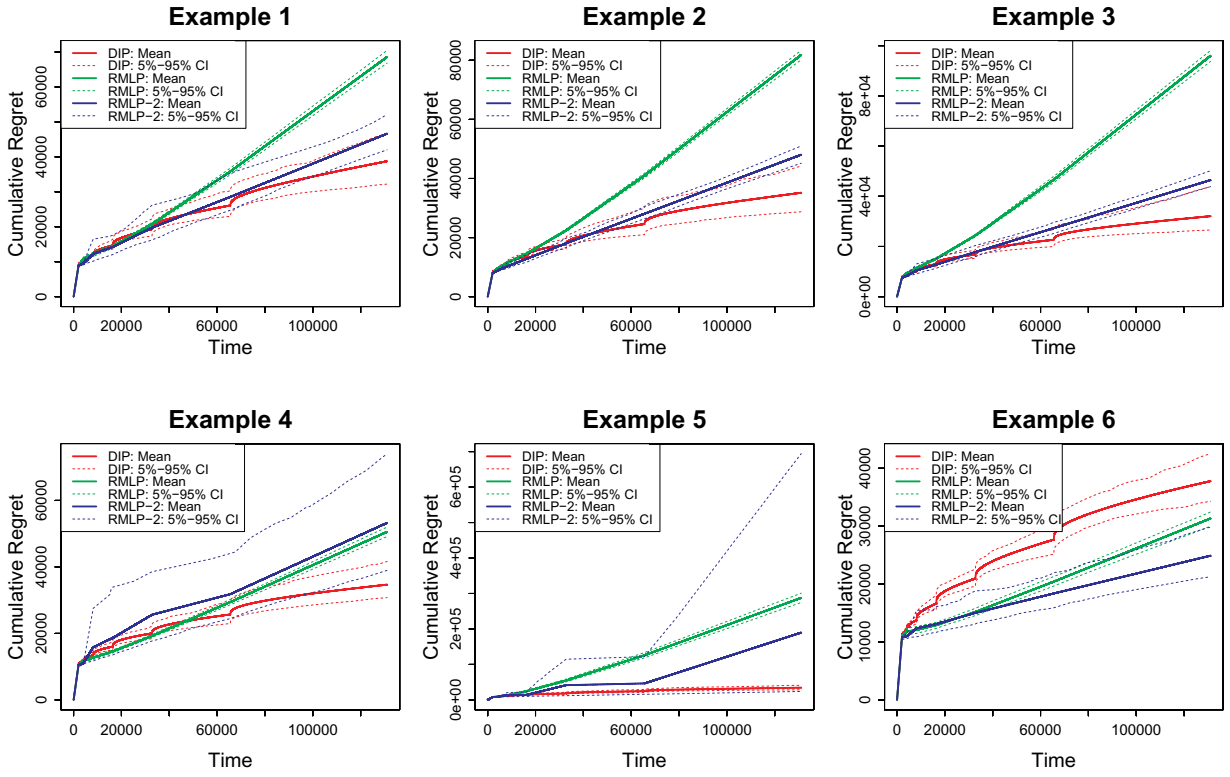
Example 7. Dimension $d_0 = 3$, $\boldsymbol{\theta}_0 = (10,10,10)^\top$, $\boldsymbol{x}_t \overset{\text{i.i.d.}}{\sim} \text{Unif}[0.3,1]^3$.

Example 8. Dimension $d_0 = 10$, $\boldsymbol{\theta}_0 = (3,\ldots,3)^\top$, $\boldsymbol{x}_t \overset{\text{i.i.d.}}{\sim} \text{Unif}[0.1,1]^{10}$.

Example 9. Dimension $d_0 = 10$, $\boldsymbol{\theta}_0 = (3,\ldots,3)^\top$, $\boldsymbol{x}_t \overset{\text{i.i.d.}}{\sim} \text{Unif}[0,1]^{10}$.

We show the log cumulative regrets of DIP, RMLP, and RMLP-2 averaged over 100 replications in Figure 6. We use log regret because the scale difference between the regrets of these three methods are large for Examples 7–9 mainly because of the unsatisfactory performance of RMLP. For all three methods, we estimate $\boldsymbol{\theta}_0$ in each of six episodes and use it for pricing in subsequent episode. Figure 7 shows boxplots of estimation errors $\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0\|_2$ for all six episodes $k = 1,\ldots,6$ and all three methods in Examples 7–9. In Examples 7 and 8, DIP outperforms RMLP-2 with more stable parameter estimations. In Example 9, the RMLP-2 is relatively stable and delivers better performance than DIP. Note that RMLP performs the worst because it specifies F as $e^x/(e^x+1)$ with the variance $\pi^2/3$, which is quite different from that of the true F. Moreover, we find that RMLP-2 sometimes obtains poor estimates and, thus, incurs large regrets. For instance, as shown in the middle plot of Figure 7 representing Example 8, there is one replication in which RMLP-2 has an estimation error more than 80 in episode 2. Then, in this replication, the regret of RMLP-2 in the subsequent episode 3 can be large because of this unsatisfactory estimate. We conjecture that the unstable estimations of RMLP-2 in Examples 7 and 8 are due to its produced singular price-covariate data in each episode. In Supplemental Section C, we provide a more detailed discussion on this phenomenon. As a comparison, DIP well-balances exploration and exploitation and sets dispersed prices at the beginning of each episode. This helps to generate a healthier data structure leading to more stable estimates.

Figure 5. Regret comparisons of DIP, RMLP, and RMLP-2 in Examples 1–6.



4.1. ℓ_1 Estimation Error Convergence

Our proven regret upper bound in Theorem 1 involves a term related to $\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0\|_1$. In our Examples 7–9 of simulations, we plot the ℓ_2 estimation errors $\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0\|_2$. In this section, we plot in Figure 8 the ℓ_1 estimation errors calculated from each episode of Example 7 to investigate its convergence rates. The left panel of Figure 8 shows a clear decaying trend starting from the second episode. In addition, in the right panel, we plot the $\log_2$ average estimation errors over the $\log_2$ number of data samples for episodes 2–6. Through a linear fit, we extract a slope of –0.354, which implies that the real decaying rate is between $-1/2$ and $-1/3$, and hence, $\alpha \in (-1/2, -1/3)$. Thus, the $\tilde{O}(T^{2/3})$ overall regret bound in our theorem can be practically achieved.

4.2. Heavy-Tailed Noise Distributions

Next, we evaluate the performance of our DIP policy on heavy-tailed noise distributions. Let Cauchy(μ, σ^2) denote the CDF of the Cauchy distribution with location parameter μ and scale parameter σ. We consider the three-

Figure 6. Log regret comparisons of DIP, RMLP, and RMLP-2 in Examples 7–9.

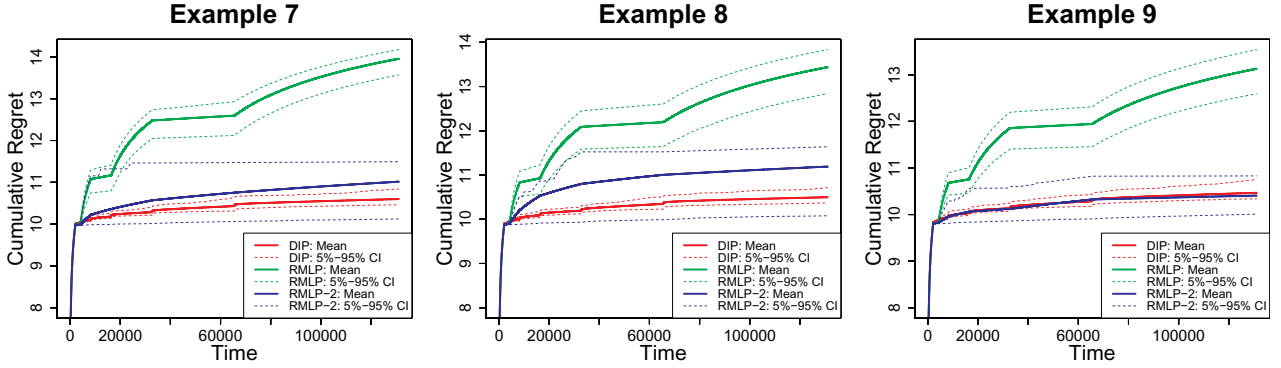
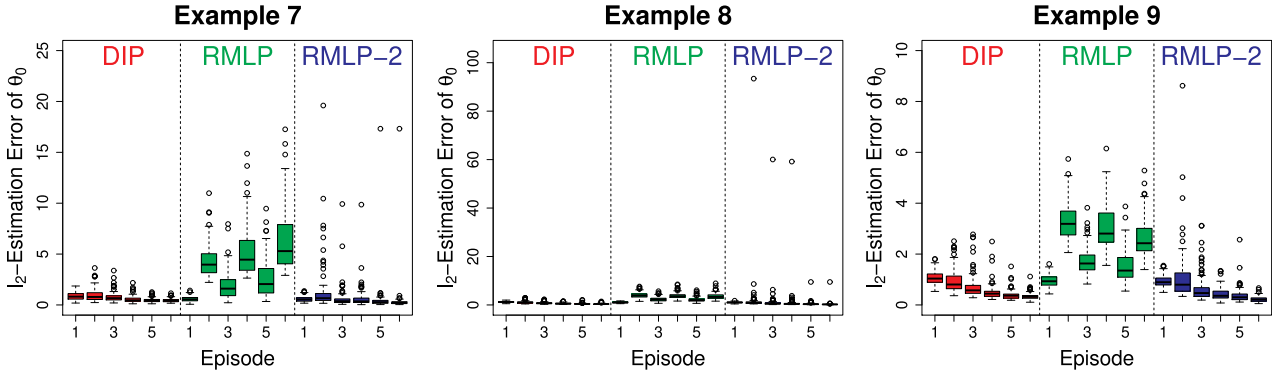


Figure 7. (Color online) Estimation errors $\|\hat{\theta}_k - \theta_0\|_2$ of DIP, RMLP, and RMLP-2 over six episodes in Examples 7–9.



dimensional covariates $\boldsymbol{x}_t \overset{\text{i.i.d.}}{\sim} \text{Unif}[0.01, 1]^3$ and set $\boldsymbol{\theta}_0 = (10, 10, 10)^\top$. The CDF F of the noise distributions are designed as follows.

Example 10. The true $F = \text{Cauchy}(0, 1)$.

Example 11. The true $F = \text{Cauchy}(0, 3)$.

Example 12. The true $F = (1/2)\text{Cauchy}(-5, 6) + (1/2)\text{Cauchy}(5, 6)$.

We repeat 100 times for each example and plot the average accumulative regret curves and their confidence bounds in Figure 9. We can see that DIP outperforms both RMLP and RMLP-2 in all these three simulation settings.

4.3. Sensitivity Tests

Our DIP policy relies on three hyperparameters $\lambda, p_{\max}$ and C. Here, λ is the regularization parameter of the regression estimation procedure, $p_{\max}$ is an upper bound of any potential optimal prices, and C is the constant in the discretization number $d = C\lceil T_0^{1/6} \rceil$. In all our simulation settings, we set $\lambda = 0.1, p_{\max} = 30$, and $C = 20$. In this section, we conduct sensitivity tests to see how their values affect the overall performance of DIP. Here, we use the simulation setting in Example 1 as an illustration. In Figure 10, we include two other values for each of the three parameters, that is, $\lambda = 0.01, 1$, $p_{\max} = 25, 35$, $C = 15, 25$. The results demonstrate that DIP is relatively robust to the values of these three parameters.

Figure 8. (Color online) ℓ_1 estimation error $\|\hat{\theta}_k - \theta_0\|_1$ of DIP in Example 7.

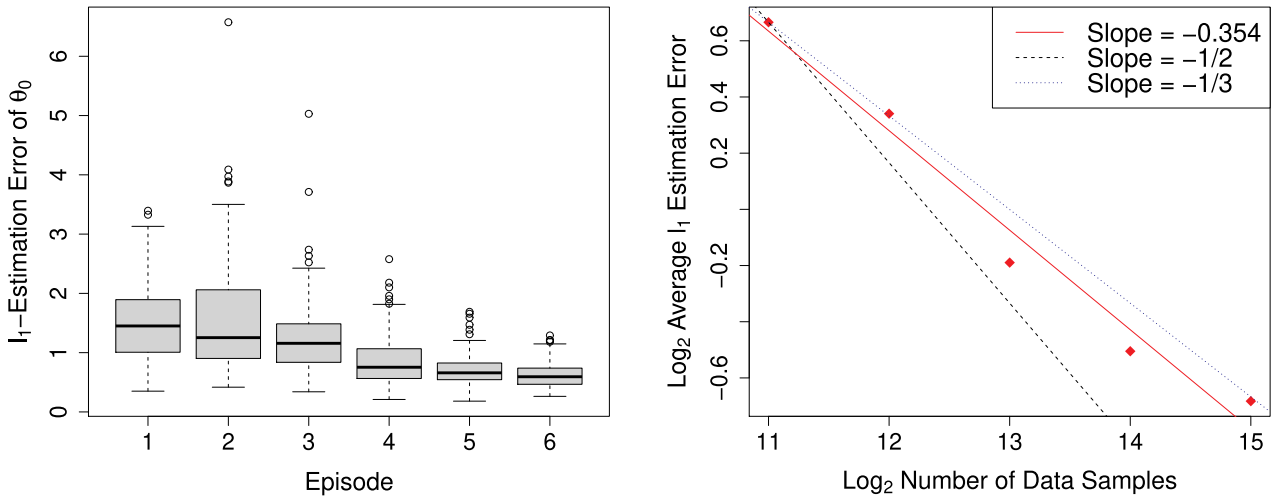
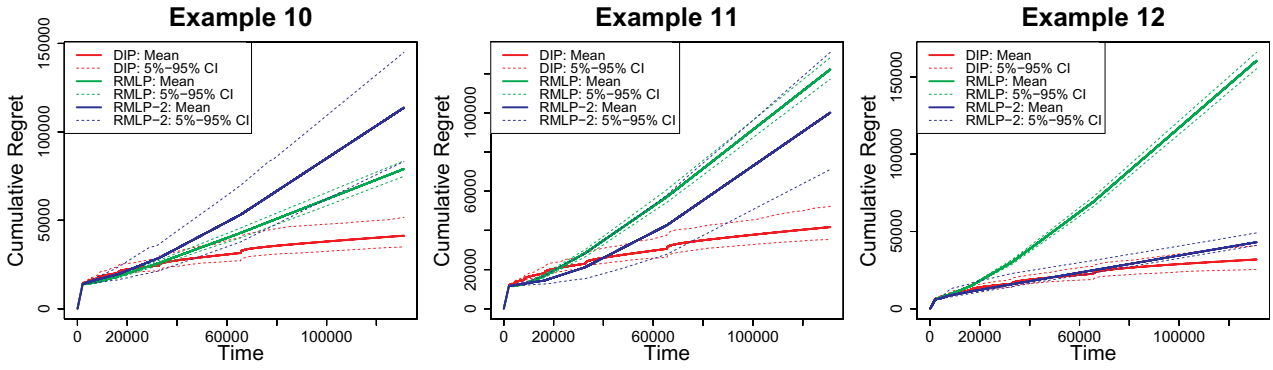


Figure 9. Regret comparisons of DIP, RMLP, and RMLP-2 in Examples 10–12.



5. Real Data Analysis

We explore the efficiency of our proposed DIP policy on a real-life auto loan data set provided by the Center for Pricing and Revenue Management at Columbia University. This data set was first studied by Phillips et al. [46] and further used by Bastani et al. [5] and Ban and Keskin [4] to evaluate different dynamic pricing algorithms.

The data set records 208,085 auto loan applications received by a major online lender in the United States from July 2002 through November 2004. For each application, we observe some loan-specific features, such as the date of application, the term and amount of loan requested, and the borrower's personal information. It also includes the monthly payment required by the lender, which can be viewed as the pricing decision. Note that it is natural to set prices according to the marketing environment, product features, and customer characteristics in online auto lending. Finally, it records whether the price was accepted by the borrower, that is, the customer's binary purchasing decision in our model.

We adopt the feature selection result used in Bastani et al. [5] and Ban and Keskin [4] and only consider the following four features: the loan amount approved, FICO score, prime rate, and the competitor's rate. We scale each feature to $[0,1]$ through dividing them by the maximum. The price p of a loan is computed as the net present value of future payment minus the loan amount, that is, $p = \text{Monthly Payment} \times \sum_{\tau=1}^{\text{Term}} (1+\text{Rate})^{-\tau} - \text{Loan Amount}$. We use \$1,000 as a basic unit and 0.12% as the rate value here, an approximate average of the monthly London interbank offered rate for the studied time period.

Note that it is impossible to obtain customers' real online responses to any dynamic pricing strategy unless it was used in the system when data were collected. Thus, we follow the off-policy learning idea used in Bastani et al. [5] and Ban and Keskin [4] to first estimate the customer choice model using the entire data set and use it as the ground truth to generate the willingness to pay of each customer given any prices. We utilize a two-step estimation procedure to estimate the unknown $\boldsymbol{\theta}_0$ and F. In particular, we use logistic regression to estimate $\boldsymbol{\theta}_0$ and then use the kernel density estimation idea to estimate F. The details of this estimation procedure are deferred to Supplemental Section D. The estimated noise PDF for the United States is shown in the left plot of Figure 11. The estimated $\hat{\boldsymbol{\theta}}_0$ and $\hat{F}$ are treated as the true parameters for the customer choice model $y_t \sim \text{Ber}(1-\hat{F}(p_t - \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_0))$. Note that these true

Figure 10. (Color online) Sensitivity tests of DIP policy with respect to $\lambda, p_{\max}$ and C.

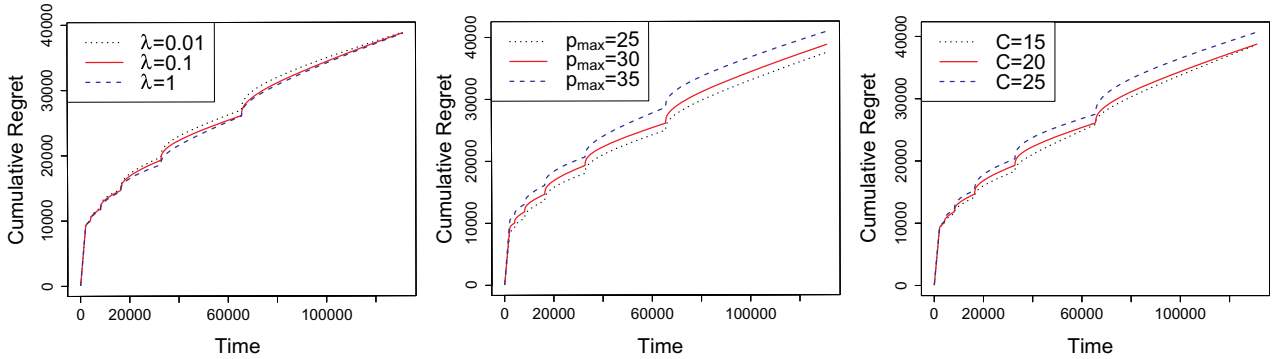
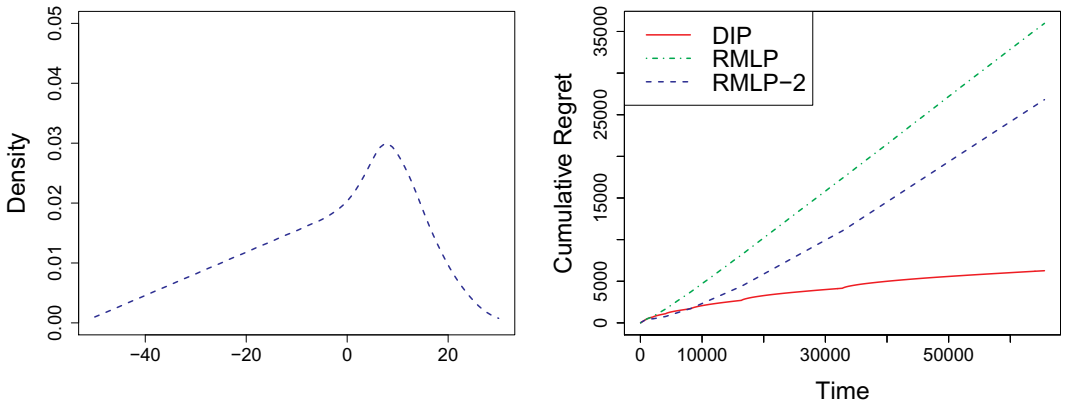


Figure 11. (Color online) The left plot shows the PDF of the noise distribution for the whole U.S. data, and the right plot shows the regret comparison of DIP, RMLP, and RMLP-2.



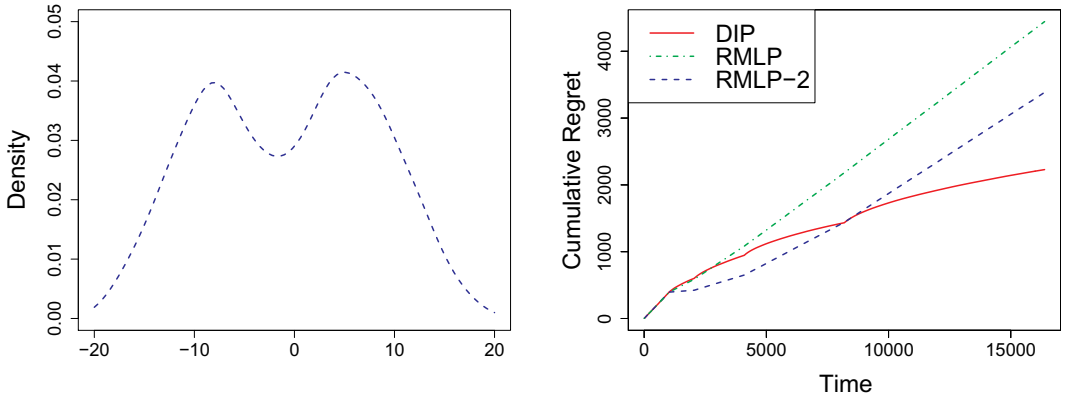
parameters are not used in any dynamic pricing algorithm, but only used to calculate the regret for any set prices and evaluate the performance of any pricing policies.

We compare DIP with RMLP-2 and RMLP policies developed in Javanmard and Nazerzadeh [32]. Because the dimension is low and the coefficients are nonsparse, we apply the RMLP-2 and RMLP policies without regularization. As required by DIP, a known upper bound $p_{\max}$ of the best prices for all applications is set as 30. We randomly sample 2^{16} applications from the total 208,085 for 50 times and apply DIP, RMLP-2, and RMLP policies to each of the 50 replications and then record the average cumulative regrets.

As shown in the right plot of Figure 11, RMLP displays the worst performance, and our proposed DIP policy outperforms RMLP-2 when the time period passes above 10^4. It enjoys more advantages as the time period grows larger. Moreover, DIP shows a clear sublinear cumulative regret, whereas RMLP-2 displays a linear pattern. This is because DIP can gradually learn the unknown distribution F. Furthermore, DIP enjoys a more accurate and stable $\boldsymbol{\theta}_0$ estimation because it invests a certain amount in price explorations and generates a more well-distributed data set. In comparison, RMLP-2 sets the prices by applying a deterministic mapping function to a linear combination of the covariates, which might yield a singular data structure that leads to unsatisfactory estimates. This phenomenon is similar to that shown in Examples 7 and 8 of the synthetic experiments.

Next, we evaluate the performance of DIP, RMLP-2, and RMLP by focusing on data in California, which has nearly 30,000 applications. We apply the same estimation procedure for $\boldsymbol{\theta}_0$ and F on the California data set to obtain the true customer choice model for California. The estimated PDF of the noise distribution for the California data are shown in the left panel of Figure 12. It has a multimodal pattern and does not satisfy the log-concave condition required by RMLP-2 and RMLP. This illustrates our motivation that the noise distribution could be complex in real applications. We record the average cumulative regrets for 50 random samplings of 2^{14} applications. As shown in the right panel of Figure 12, DIP again achieves a sublinear regret and outperforms that of RMLP-2 eventually. Similar to the previous case, RMLP performs worse than DIP and RMLP-2.

Figure 12. (Color online) The left plot shows the PDF of the noise distribution for the California data, and the right plot shows the regret comparison of DIP, RMLP, and RMLP-2.



6. Conclusion

In this paper, we consider a customer choice model generated by a linear valuation function with the unknown coefficient parameter and unknown noise distribution. A new pricing policy DIP is proposed to tackle this problem through simultaneously learning both the unknown parameter and the unknown distribution. In theory, we show that, even when the noise distribution is unknown, our DIP policy is still able to achieve a sublinear regret bound. We apply DIP on various synthetic data sets and a real online auto loan data set and demonstrate its superior performance when compared with state-of-the-art pricing algorithms.

There are a few interesting future directions. In this paper, we focus on nonsparse coefficients with an unknown noise distribution. It would be interesting to extend our policy to the high-dimensional setting with a sparse linear choice model. We can also extend the linear choice model to a more flexible semiparametric model (Bickel et al. [9]) to allow both a parametric component and a nonparametric component on the covariates. Furthermore, it would be interesting to incorporate the considerations of fairness and welfare (Kallus and Zhou [34]) into our dynamic pricing regime.

Acknowledgments

The authors are indebted to the editor, the associate editor, and two referees, whose helpful comments and suggestions led to a much improved presentation. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not reflect the views of the National Science Foundation.

References

[1] Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. *Adv. Neural Inform. Processing Systems* 24:2312–2320.
[2] Agrawal S, Goyal N (2013) Thompson sampling for contextual bandits with linear payoffs. *Proc. Internat. Conf. Machine Learn.*, 127–135.
[3] Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Machine Learn.* 47(2):235–256.
[4] Ban GY, Keskin NB (2021) Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Sci.* 67(9):5549–5568.
[5] Bastani H, Simchi-Levi D, Zhu R (2022) Meta dynamic pricing: Transfer learning across experiments. *Management Sci.* 68(3):1865–1881.
[6] Besbes O, Zeevi A (2009) Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Oper. Res.* 57(6):1407–1420.
[7] Besbes O, Zeevi A (2011) On the minimax complexity of pricing in a changing environment. *Oper. Res.* 59(1):66–79.
[8] Besbes O, Zeevi A (2015) On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Sci.* 61(4):723–739.
[9] Bickel PJ, Klaassen CA, Ritov Y, Wellner JA (1998) *Efficient and Adaptive Estimation for Semiparametric Models* (Springer, New York).
[10] Broder J, Rusmevichientong P (2012) Dynamic pricing under a general parametric choice model. *Oper. Res.* 60(4):965–980.
[11] Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations Trends Machine Learn.* 5(1):1–122.
[12] Cesa-Bianchi N, Cesari T, Perchet V (2019) Dynamic pricing with finitely many unknown valuations. *Proc. 30th Internat. Conf. Algorithmic Learn. Theory*, 247–273.
[13] Chen N, Gallego G (2021) Nonparametric pricing analytics with customer covariates. *Oper. Res.* 69(3):974–984.
[14] Chen N, Gallego G (2022) A primal–dual learning algorithm for personalized dynamic pricing with an inventory constraint. *Math. Oper. Res.* 47(4):2585–2613.
[15] Chen Y, Wen Z, Xie Y (2019) Dynamic pricing in an evolving and unknown marketplace. Preprint, submitted June 6, https://dx.doi.org/10.2139/ssrn.3382957.
[16] Chen X, Owen Z, Pixton C, Simchi-Levi D (2022) A statistical learning approach to personalization in revenue management. *Management Sci.* 68(3):1923–1937.
[17] Cheung WC, Simchi-Levi D, Wang H (2017) Dynamic pricing and demand learning with limited price experimentation. *Oper. Res.* 65(6):1722–1731.
[18] Cheung WC, Simchi-Levi D, Zhu R (2022) Hedging the drift: Learning to optimize under nonstationarity. *Management Sci.* 68(3):1696–1713.
[19] Chu W, Li L, Reyzin L, Schapire R (2011) Contextual bandits with linear payoff functions. *Proc. Internat. Conf. Artificial Intelligence Statistics*, 208–214.
[20] Cohen MC, Lobel I, Paes Leme R (2020) Feature-based dynamic pricing. *Management Sci.* 66(11):4921–4943.
[21] den Boer AV (2015) Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys Oper. Res. Management Sci.* 20(1):1–18.
[22] den Boer AV (2015) Tracking the market: Dynamic pricing and learning in a changing environment. *Eur. J. Oper. Res.* 247(3):914–927.
[23] den Boer AV, Keskin NB (2020) Discontinuous demand functions: Estimation and pricing. *Management Sci.* 66(10):4516–4534.
[24] den Boer AV, Zwart B (2014) Simultaneously learning and optimizing using controlled variance pricing. *Management Sci.* 60(3):770–783.
[25] Fan J, Guo Y, Yu M (2022) Policy optimization using semiparametric models for dynamic pricing. *J. Amer. Statist. Assoc.* 1–29.
[26] Foster D, Rakhlin A (2020) Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *Proc. Internat. Conf. Machine Learn.*, 3199–3210.
[27] Foster DJ, Gentile C, Mohri M, Zimmert J (2020) Adapting to misspecification in contextual bandits. *Adv. Neural Inform. Processing Systems* 33:11478–11489.

[28] Golrezaei N, Jaillet P, Liang JCN (2019) Incentive-aware contextual pricing with non-parametric market noise. Preprint, submitted November 8, https://arxiv.org/abs/1911.03508.
[29] Golrezaei N, Javanmard A, Mirrokni V (2021) Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Oper. Res.* 69(1):297–314.
[30] Huang J, Mani A, Wang Z (2022) The value of price discrimination in large social networks. *Management Sci.* 68(6):4454–4477.
[31] Javanmard A (2017) Perishability of data: Dynamic pricing under varying-coefficient models. *J. Machine Learn. Res.* 18(1):1714–1744.
[32] Javanmard A, Nazerzadeh H (2019) Dynamic pricing in high-dimensions. *J. Machine Learn. Res.* 20(1):315–363.
[33] Javanmard A, Nazerzadeh H, Shao S (2020) Multi-product dynamic pricing in high-dimensions with heterogeneous price sensitivity. *Proc. IEEE Internat. Sympos. Inform. Theory*, 2652–2657.
[34] Kallus N, Zhou A (2021) Fairness, welfare, and equity in personalized pricing. *Proc. Conf. Fairness Accountability Transparency*, 296–314.
[35] Keskin NB, Zeevi A (2014) Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Oper. Res.* 62(5):1142–1167.
[36] Keskin NB, Zeevi A (2017) Chasing demand: Learning and earning in a changing environment. *Math. Oper. Res.* 42(2):277–307.
[37] Kleinberg R, Leighton T (2003) The value of knowing a demand curve: Bounds on regret for online posted-price auctions. *Proc. IEEE Sympos. Foundations Comput. Sci.* (IEEE, Piscataway, NJ), 594–605.
[38] Lattimore T, Szepesvári C (2020) *Bandit Algorithms* (Cambridge University Press, Cambridge, UK).
[39] Lattimore T, Szepesvari C, Weisz G (2020) Learning with good feature representations in bandits and in RL with a generative model. *Proc. Internat. Conf. Machine Learn.*, 5662–5670.
[40] Mao J, Leme R, Schneider J (2018) Contextual pricing for Lipschitz buyers. *Adv. Neural Inform. Processing Systems* 31:5643–5651.
[41] Misra K, Schwartz EM, Abernethy J (2019) Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Sci.* 38(2):226–252.
[42] Mueller JW, Syrgkanis V, Taddy M (2019) Low-rank bandit methods for high-dimensional dynamic pricing. *Adv. Neural Inform. Processing Systems* 32:15442–15452.
[43] Nambiar M, Simchi-Levi D, Wang H (2019) Dynamic learning and pricing with model misspecification. *Management Sci.* 65(11):4980–5000.
[44] Pacchiano A, Phan M, Abbasi Yadkori Y, Rao A, Zimmert J, Lattimore T, Szepesvari C (2020) Model selection in contextual stochastic bandit problems. *Adv. Neural Inform. Processing Systems* 33:10328–10337.
[45] Perchet V, Rigollet P (2013) The multi-armed bandit problem with covariates. *Ann. Statist.* 41(2):693–721.
[46] Phillips R, Şimşek AS, Van Ryzin G (2015) The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Sci.* 61(8):1741–1759.
[47] Qiang S, Bayati M (2016) Dynamic pricing with demand covariates. Preprint, submitted April 25, https://arxiv.org/abs/1604.07463.
[48] Russac Y, Vernade C, Cappé O (2019) Weighted linear bandits for non-stationary environments. *Adv. Neural Inform. Processing Systems* 32:12017–12026.
[49] Shah V, Johari R, Blanchet J (2019) Semi-parametric dynamic contextual pricing. *Adv. Neural Inform. Processing Systems* 32:2363–2373.
[50] Wang J, Shen X, Liu Y (2008) Probability estimation for large-margin classifiers. *Biometrika.* 95(1):149–167.
[51] Wang Y, Chen B, Simchi-Levi D (2021) Multimodal dynamic pricing. *Management Sci.* 67(10):6136–6152.
[52] Wang Z, Deng S, Ye Y (2014) Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Oper. Res.* 62(2):318–331.
[53] Wang Y, Chen X, Chang X, Ge D (2021) Uncertainty quantification for demand prediction in contextual dynamic pricing. *Production Oper. Management* 30(6):1703–1717.
[54] Xu J, Wang YX (2021) Logarithmic regret in feature-based dynamic pricing. *Adv. Neural Inform. Processing Systems* 34:13898–13910.
[55] Xu J, Wang YX (2022) Toward agnostic feature-based dynamic pricing: Linear policies vs linear valuation with unknown noise. *Proc. Internat. Conf. Artificial Intelligence Statistics*, 9643–9662.
[56] Zhao P, Zhang L, Jiang Y, Zhou ZH (2020) A simple approach for non-stationary linear bandits. *Proc. Internat. Conf. Artificial Intelligence Statistics*, 746–755.