# Distribution-free Contextual Dynamic Pricing

Yiyun Luo*, Will Wei Sun†, and Yufeng Liu‡

**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 historical binary purchase feedbacks. 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 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, where 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 sub-linear regret bound in the considered pricing problem. Finally, we demonstrate the superior performance of our policy on simulations and a real-life auto-loan dataset.

**Keywords:** Classification; Dynamic Pricing; Linear Bandits; Regret Analysis

*PhD Student, Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill. Email: yiyun851@live.unc.edu.

†Assistant Professor, Krannert School of Management, Purdue University. Email: sun244@purdue.edu.

‡Professor, Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill. Email: yfliu@email.unc.edu.

# 1 Introduction

Contextual dynamic pricing aims to design an online pricing policy adaptive to product features, customer characteristics, and marketing environment (Huang et al., 2021). It has been widely used in industries such as hospitality, tourism, entertainment, retail, electricity, and public transportation (den Boer, 2015a). A successful dynamic pricing algorithm involves both pricing and learning to maximize the revenues. Upon receiving sequential customer responses, the algorithm continuously updates its knowledge on the 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, i.e., 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 has been studied extensively in dynamic pricing (Besbes and Zeevi, 2009; Keskin and Zeevi, 2014; Cheung et al., 2017; Chen et al., 2019; Cesa-Bianchi et al., 2019; den Boer and Keskin, 2020). In this case, one critical task is to balance the tradeoff between exploration and exploitation, where 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 (Javanmard and Nazerzadeh, 2019; Golrezaei et al., 2019). Given the contextual covariate $x$, e.g., product features, customer characteristics, and marketing environment, the customer's valuation $v(x)$ for the product is $v(x) = x^\top \theta_0 + z$. Here, the first component represents the linear effect of the covariates $x$ with an unknown parameter $\theta_0$, 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 $v(x)$ 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 (2019) assumes a known $F$ for their RMLP method and considers $F$ to belong to a log-concave family for their RMLP-2

policy. In spite that knowing $F$ 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 dataset studied in Section 5, the estimated Probability Distribution Functions (PDF) of the noise term in four states are clearly not log-concave, as shown in Figure 1.

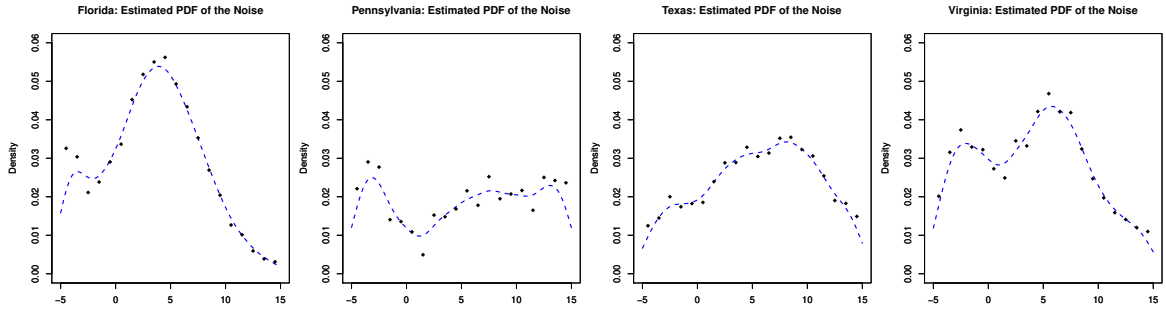


Figure 1: Estimated noise PDFs for four states in our auto loan real application.

In this paper, we propose a DIstribution-free Pricing (DIP) policy to tackle the contextual dynamic pricing problem with unknown $\theta_0$ and unknown $F$. DIP employs a doubling trick (Lattimore and Szepesvári, 2020) 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 $\theta_0$ estimation into a classification problem in which no prior knowledge of $F$ is required, our DIP policy adopts the logistic regression to obtain a $\theta_0$ estimate 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 as an extension of the classic linear bandit (Abbasi-Yadkori et al., 2011; Chu et al., 2011; Agrawal and Goyal, 2013), and is also of independent interests. Interestingly, the "perturbation level" of the translated PLB can be specified as proportional to the $\ell_1$ error of the given $\theta_0$ 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 conduct extensive theoretical investigations of our DIP policy. The regret, as the expected revenues loss 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 a sub-linear term and a linear term is analogous to the regret in misspecified linear bandits (Lattimore et al., 2020; Pacchiano et al., 2020; Foster et al., 2020). However, by utilizing the unique structure of our PLB setting, we improve the regret bounds of existing misspecified linear bandits (Lattimore et al., 2020; Pacchiano et al., 2020; Foster et al., 2020) by saving a $\sqrt{d}$ on the linear term. See our Remark 2 for more discussion. Importantly, we also show that the linear dependence of $T_0$ is unavoidable by establishing a matching lower bound for our PLB problem. 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 $\tilde{\mathcal{O}}(T^{2/3})$ sub-linear term and an extra term related to the estimation error for $\theta_0$. The latter term is dominated by the sub-linear term in a broad range of scenarios, which is well supported by our experiments. In summary, our sub-linear regret upper bound implies that the average regret per time period vanishes as the time horizon tends to infinity.

Finally, we demonstrate the superior performance of our policy on extensive simulations and a real-life auto-loan dataset by comparing our DIP policy to RMLP and RMLP-2 (Javanmard and Nazerzadeh, 2019). Due to 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 dataset, our DIP policy is shown to largely improve the cumulative 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 cumulative regret over the considered time horizon. Such an improvement keeps increasing when the total time horizon increases. See Figure 8 and Section 5 for more details.

## 1.1 Related Work

**Non-contextual dynamic pricing.** For non-contextual dynamic pricing without covariates, Besbes and Zeevi (2009); Wang et al. (2014); Besbes and Zeevi (2015); Chen and Gallego (2018) designed policies to handle a nonparametric model while Besbes and Zeevi (2009); Broder and Rusmevichientong (2012); den Boer and Zwart (2014); Keskin and Zeevi (2014) considered parametric models. Furthermore, Besbes and Zeevi (2011); den Boer (2015b); Keskin and Zeevi (2017) investigated the time-varying unknown demand setting. In addition, the Upper Confidence Bound (UCB) idea (Auer et al., 2002; Abbasi-Yadkori et al., 2011) has been used in different non-contextual instances (Kleinberg and Leighton, 2003; Misra et al., 2019; Wang et al., 2021). However, all these approaches do not incorporate the covariates into pricing policy. Therefore, our model and technical tools are fundamentally different.

**Contextual dynamic pricing.** Dynamic pricing with covariates has garnered significant interest among researchers. As Mueller et al. (2018); Javanmard et al. (2020); Chen et al. (2021) focused on the multi-product setting, most of contextual dynamic pricing literature (Qiang and Bayati, 2016; Javanmard, 2017; Mao et al., 2018; Nambiar et al., 2019; Bastani et al., 2019; Cohen et al., 2020; Wang et al., 2020; Ban and Keskin, 2020) considered a single product at each time. It is interesting that the unknown distribution of the random noise in our linear valuation model transfers to some unknown nonparametric link function in the demand model, which greatly extends the existing work which assumes a known (generalized) linear link function in the demand model (Ban and Keskin, 2020). Moreover, Javanmard and Nazerzadeh (2019); Golrezaei et al. (2019, 2021) also considered the linear valuation model as we do. Similar to us, Golrezaei et al. (2019) assumed both the unknown linear effect and noise distribution and thus faced the same challenge of error propagation. They adopted a second price auction mechanism with multiple buyers at each time. One main difference lies in the feedback structure. Namely, they assumed a "full information" setting that the seller observed all bids and valuations from multiple buyers while we considered a "bandit" setting that the seller only observed one single buyer's binary purchasing decision. In Javanmard and Nazerzadeh (2019), their proposed RMLP assumed a known market noise distribution while RMLP-2 assumed 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. (2021) also established a $\tilde{\mathcal{O}}(T^{2/3})$ 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 to the true optimal policy. On the other hand, Shah et al. (2019); Chen and Gallego (2021) shared similar nonparametric ingredients in the unknown demand function as ours. Specifically, Chen and Gallego (2021) considered a general Lipschitz demand and proposed a pricing policy based on adaptive binning of the covariate space (Perchet et al., 2013) with a regret of $\tilde{\mathcal{O}}(T^{(2+d_0)/(4+d_0)})$, 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. (2019) adopted a log-linear valuation model to handle the unknown nonparametric noise in their semi-parametric model. Their method heavily relies on the special structure of the log-linear valuation model, whose optimal price has nice separable effects of the unknown linear structure and unknown noise distribution. Hence their approach is not applicable to our pricing model where these two unknown parts entangle with each other. Therefore, techniques used in Shah et al. (2019); Chen and Gallego (2021) for handling nonparametric components in the demand function are very different from the newly-proposed PLB framework of our DIP policy.

**Bandit algorithms.** Our pricing policy is also related to bandit algorithms (Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvári, 2020; Foster and Rakhlin, 2020) which address the balance between exploration and exploitation. In particular, our perturbed linear bandit is related to misspecified linear bandits (Lattimore et al., 2020; Pacchiano et al., 2020; Foster et al., 2020) and non-stationary linear bandits (Cheung et al., 2018; Russac et al., 2019; Zhao et al., 2020). 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 to a direct application of much complex existing algorithms for misspecified or non-stationary linear bandits. See Remarks 2 and 3 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 $\beta \in \mathbb{R}^d$, let $\|\beta\|_\infty = \max_j |\beta_j|$ and $\|\beta\|_1 = \sum_{j=1}^d |\beta_j|$ denote its max norm and $\ell_1$ norm, respectively. For two sequences $a_n, b_n$, we say $a_n = \mathcal{O}(b_n)$ if $a_n \leq Cb_n$ for some positive constant $C$, $a_n = \tilde{\mathcal{O}}(b_n)$ if $a_n = \mathcal{O}(b_n)$ that ignores a logarithm term, 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 datasets and in Section 5, we apply DIP to a real-life auto loan dataset. We conclude our work along with some future directions in Section 6. Most technical proofs are collected in the Supplementary Material.

## 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 $t \in [T] = \{1, \ldots, T\}$, and the seller observes a covariate $x_t \in \mathcal{X} \subseteq \mathbb{R}^{d_0}$ representing the product features and customer characteristics. Similar to Javanmard and Nazerzadeh (2019); Golrezaei et al. (2019); Shah et al. (2019); Chen and Gallego (2021), we assume $||x_t||_\infty \leq 1, \forall x_t \in \mathcal{X}$. Given $x_t$, the customer's valuation of the product $v_t = v(x_t) = x_t^\top \theta_0 + z_t$ is a sum of a linear function of $x_t$ and a market noise $z_t$. We assume $\{z_t\}_{t\in[T]}$ are drawn 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 $y_t = 1_{\{v_t \geq p_t\}}$ denote whether the customer buys the

product. By the aforementioned sales mechanism, it follows

$$y_t = \begin{cases} 1 & \text{if } v_t \geq p_t, \text{ with probability } 1 - F(p_t - x_t^\top \theta_0); \\ 0 & \text{if } v_t < p_t, \text{ with probability } F(p_t - x_t^\top \theta_0), \end{cases}$$

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

Given the above customer choice model and the covariate $x$, the expected reward of a setting price $p$ is $p(1 - F(p - x^\top \theta_0))$. We define the optimal price $p^*(x)$ as that maximizing $p(1 - F(p - x^\top \theta_0))$, which is an implicit function of the covariate and dependent on both the unknown $\theta_0$ and $F$. By dynamically setting prices and observing binary feedbacks, we collect instant revenues and meanwhile gather more information to estimate $\theta_0, F$ and $p^*(x)$. An important feature of this process is the tradeoff between exploration and exploitation where we shall 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

$$p_t^* = p^*(x_t) = \arg\max_{p>0} p(1 - F(p - x_t^\top \theta_0))$$

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 to the optimal price $p_t^*$, i.e.,

$$r_t = p_t^*(1 - F(p_t^* - x_t^\top \theta_0)) - p_t(1 - F(p_t - x_t^\top \theta_0)). \tag{1}$$

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 $x_t$ at time $t$, by utilizing all historical data $\{(x_s, p_s, y_s), s = 1, \ldots, t-1\}$, 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 the Inner Algorithm A and the Inner Algorithm B. The Inner Algorithm A is designed for estimating $\theta_0$, and the 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 tradeoff 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, 2020) 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 of the current one until the horizon ends. The number of episodes $n = n(T, \alpha_1, \alpha_2)$ 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 $\alpha_1, \alpha_2$. Figure 2 demonstrates the case when the total time horizon is cut into 5 episodes via the doubling trick.

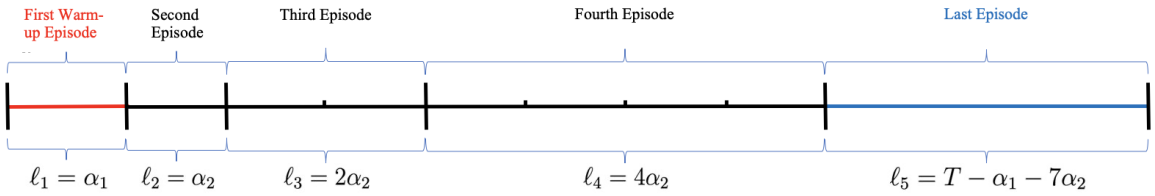


Figure 2: An illustration of cutting total time horizon utilizing the doubling trick.

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 $\theta_0$ and Inner Algorithm B to set prices. Specifically, Inner Algorithm A uses all data from episode $k-1$ to obtain an estimate $\hat{\theta}_{k-1}$ of $\theta_0$; then Inner Algorithm B takes $\hat{\theta}_{k-1}$ 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 $\theta_0$ 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

**Input: (arrives over time)** covariates $\{x_t\}_{t\in[T]}$
Denote the episodes yielded by the doubling trick as $\mathcal{E}_1, \ldots, \mathcal{E}_n$.
**For** $t \in \mathcal{E}_1$ **do**
  Set a price $p_t$ randomly from $(0, p_{\max})$ and receive a binary response $y_t$.
Apply Inner Algorithm A on $\{(x_t, p_t, y_t)\}_{t\in\mathcal{E}_1}$ to obtain an estimate $\hat{\theta}_1$ of $\theta_0$.
**For** episode k=$2, 3, \ldots, n$ **do**
  With input $\hat{\theta}_{k-1}$ as the estimate of $\theta_0$, apply Inner Algorithm B on $\mathcal{E}_k$ to sequentially set a price $p_t$ and receive a binary response $y_t$ for all $t \in \mathcal{E}_k$.
  With input $\{(x_t, p_t, y_t)\}_{t\in\mathcal{E}_k}$, apply Inner Algorithm A on this dataset to update an estimate $\hat{\theta}_k$ of $\theta_0$.

### 2.2.2 Inner Algorithm A

Now we introduce the Inner Algorithm A designed for estimating $\theta_0$. It uses all data $(x_t, p_t, y_t)$ from the $(k-1)$-th episode to obtain an estimate $\hat{\theta}_{k-1}$ for future pricing in the $k$-th episode. For simplicity, we introduce its generic version with $[T_0] = \{1, \ldots, T_0\}$ representing the $(k-1)$-th episode horizon. Since $y_t$ is binary and invoked by $x_t, p_t$ through $\mathbb{P}(y_t = 1) = 1 - F(p_t - x_t^\top\theta_0)$, we obtain

$$\begin{cases} \mathbb{P}(y_t = 1) > \frac{1}{2}, \text{ if } F^{-1}(\frac{1}{2}) + x_t^\top\theta_0 - p_t > 0; \\ \mathbb{P}(y_t = 1) = \frac{1}{2}, \text{ if } F^{-1}(\frac{1}{2}) + x_t^\top\theta_0 - p_t = 0; \\ \mathbb{P}(y_t = 1) < \frac{1}{2}, \text{ if } F^{-1}(\frac{1}{2}) + x_t^\top\theta_0 - p_t < 0. \end{cases}$$

Therefore, we can form a classification problem with responses $y_t$ and covariates $(1, x_t^\top, p_t)^\top$ for $t \in [T_0]$. It admits a Bayes decision boundary $\{u : (F^{-1}(\frac{1}{2}), \theta_0^\top, -1)u = 0\}$ which involves the unknown parameter $\theta_0$. Thus we can estimate the linear decision boundary and extract an estimate of $\theta_0$ by applying a linear classification method. In this paper, we use logistic regression, which yields an estimate $(\hat{c}, \hat{\beta}^\top, \hat{b})$ of $(F^{-1}(\frac{1}{2}), \theta_0^\top, -1)$ up to a constant factor. Thus $-\frac{\hat{\beta}}{\hat{b}}$ is a natural estimate of $\theta_0$. Similar to Javanmard and Nazerzadeh (2019), we assume $||\theta_0||_1$ is upper bounded by a known constant $W$. By projecting $-\frac{\hat{\beta}}{\hat{b}}$ onto the $\ell_1$-ball $\Theta = \{\theta \in \mathbb{R}^{d_0} : ||\theta||_1 \le W\}$, we can obtain our final estimate denoted as $\hat{\theta} = \text{Proj}_\Theta(-\frac{\hat{\beta}}{\hat{b}})$. Such a projection has a closed-form solution as $\text{Proj}_\Theta(-\frac{\hat{\beta}}{\hat{b}}) = \mathcal{T}_{\rho_{\min}}(-\frac{\hat{\beta}}{\hat{b}})$, where $\mathcal{T}_\rho(v) = \text{sgn}(v)(|v| - \rho)_+$ is the soft-thresholding operator and $\rho_{\min} = \min\{\rho : ||\mathcal{T}_\rho(-\frac{\hat{\beta}}{\hat{b}})||_1 \le W\}$. Here

the assumption of constant $W$ is purely for theoretical purpose 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:** $\{(x_t, p_t, y_t)\}_{t\in[T_0]}$, $W$

2: Use logistic regression to obtain the minimizer

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

3: Estimate $\theta_0$ by $\hat{\theta} = \text{Proj}_\Theta(-\frac{\hat{\beta}}{\hat{b}})$, where $\Theta = \{\theta \in \mathbb{R}^{d_0} : ||\theta||_1 \leq W\}$.

Under the same assumption of a known upper bound $W$ of $||\theta_0||_1$, RMLP and RMLP-2 in Javanmard and Nazerzadeh (2019) estimated $\theta_0$ via the maximum likelihood type of method by assuming some knowledge on $F$. In comparison, our approach achieves robust $\theta_0$ 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, e.g., large-margin classifiers (Wang et al., 2008). 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 $\hat{\theta}_{k-1}$ 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 $\hat{\theta}$ to represent $\hat{\theta}_{k-1}$ 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 $p(1-F(p-x_t^\top\theta_0))$ plays a critical role in deciding the best price at time $t$. Given the current estimate $\hat{\theta}$, we would need to evaluate $\{p(1 - F(p - x_t^\top\hat{\theta}))\}$ 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 was also used in Javanmard and Nazerzadeh (2019); Chen and Gallego (2021). By the condition $||x_t||_\infty \leq 1$, we have $p - x_t^\top\hat{\theta} \in G(\hat{\theta}) = [-||\hat{\theta}||_1, p_{\max} + ||\hat{\theta}||_1]$. Therefore the evaluation of the expected reward is reduced to evaluate $1 - F$ on $G(\hat{\theta})$. When $F$ is Lipschitz

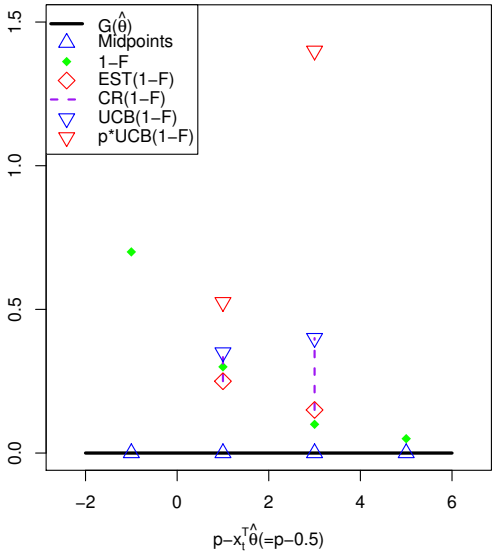


Figure 3: An illustration of Inner Algorithm B via Example 1.

continuous and no other global smoothness is assumed, it is sufficient to evaluate $1-F$ on several well-chosen discrete points in $G(\hat{\theta})$ to leverage the finite data for better pricing. In this paper, we utilize the discretization idea (Kleinberg and Leighton, 2003; Weed et al., 2016) to cut $G(\hat{\theta})$ into $d$ same-length subintervals with the set of their midpoints $\mathcal{M} = \{m_1, \ldots, m_d\}$. Here $d$ is a parameter that possibly depends on the horizon length $T_0$. When $T_0$ is large, it would be reasonable to set a bigger $d$ 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 $\mathcal{M}$.

**Example 1** (**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 will be introduced in our pricing policy for this specific example are displayed in Figure 3. Consider a two-dimensional covariate* $x_t = (0.3, 0.2)^\top$ *at time* $t$. *Assume we have an estimation* $\hat{\theta} = (1,1)^\top$ *and* $p_{\max} = 4$. *Then the interval for discretization is* $G(\hat{\theta}) = [-||\hat{\theta}||_1, p_{\max} + ||\hat{\theta}||_1] = [-2, 6]$ *represented by the black solid line in Figure 3. If* $d = 4$, *we discretize* $G(\hat{\theta})$ *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 will 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 $\mathcal{S}_t = \{m_j + x_t^\top \hat{\theta} | j \in [d], m_j + x_t^\top \hat{\theta} \in (0, p_{\max})\}$. The key feature for any price $p \in \mathcal{S}_t$ is that $p - x_t^\top \hat{\theta}$ exactly equals to 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 $p = m_j + x_t^\top \hat{\theta} \in \mathcal{S}_t$, we can leverage our current knowledge of $1 - F(m_j)$ to obtain an estimate of its expected reward $p(1 - F(p - x_t^\top \theta_0))$ as $p(1-F(p-x_t^\top \hat{\theta})) = p(1-F(m_j))$. 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 $p_t = m_j + x_t^\top \hat{\theta}$ from $\mathcal{S}_t$, we observe a binary response $y_t \sim \text{Ber}(1 - F(m_j + x_t^\top \hat{\theta} - x_t^\top \theta_0)) \approx \text{Ber}(1 - F(m_j))$ 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 $p_t = m_j + x_t^\top \hat{\theta}$. Then pulling arm $j$ yields more knowledge for $1 - F$ on $m_j$. Thus we define the available arm set at time $t$ as $\mathcal{B}_t = \{j \in [d] : \exists p \in \mathcal{S}_t \text{ such that } p = m_j + x_t^\top \hat{\theta}\}$, which varies over time as $x_t^\top \hat{\theta}$ changes over time.

**Example 1** (**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* $x_t^\top \hat{\theta} = 0.5$*, we obtain* $\mathcal{S}_t = \{m_j + x_t^\top \hat{\theta} | m_j + x_t^\top \hat{\theta} \in (0, p_{\max})\} = \{m_2 + 0.5, m_3 + 0.5\} = \{1.5, 3.5\}$ *since* $m_1 + x_t^\top \hat{\theta} = -0.5$ *and* $m_4 + x_t^\top \hat{\theta} = 5.5$ *are out of the range* $(0, p_{\max})$*. In this case, the arm set at time* $t$ *is* $\mathcal{B}_t = \{j \in [d] : \exists p \in \mathcal{S}_t \text{ such that } p = m_j + x_t^\top \hat{\theta}\} = \{2, 3\}$*.*

Restricted on $\mathcal{S}_t$, there is a clear tradeoff 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 while a pure exploitation may continuously pull suboptimal arms due to 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, 2020) to construct an upper confidence bound (UCB), 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 subsection as they emerge naturally from the perturbed linear bandit formulation of our single-episode pricing problem. Then we select $p_t = m_j + x_t^\top \hat{\theta} \in \mathcal{S}_t$ with the largest optimism estimation $p_t \mathrm{UCB}_t(1-F(m_j))$, where $\mathrm{UCB}_t(1-F(m_j)) = \mathrm{EST}_t(1-F(m_j)) + \mathrm{CR}_t(1-F(m_j))$ is an optimism estimation of $1-F(m_j)$. This optimism estimation addresses the exploration-exploitation tradeoff since a large UCB could result in either exploring a less-pulled arm with a large CR or exploiting an optimal arm with a large mean estimation.

**Example 1 (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 $m_2 + x_t^\top \hat{\theta}$ and $m_3 + x_t^\top \hat{\theta}$. To emphasize this, in Figure 3, we only show $\{EST_t(1 - F(m_j))\}_{j=2,3}$ (red hollow diamonds) and $\{CR_t(1 - F(m_j))\}_{j=2,3}$ (lengths of purple dashed line) at two midpoints $m_2 = 1$ and $m_3 = 3$. Summing them up leads to the optimism estimations $\{UCB_t(1-F(m_j))\}_{j=2,3}$ represented by blue hollow inverted triangles. Multiplying them by their corresponding prices $m_2 + x_t^\top \hat{\theta} = 1.5$ and $m_3 + x_t^\top \hat{\theta} = 3.5$, 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 will set the price $p_t = 3.5$, i.e., $m_3 + x_t^\top \hat{\theta}$, since $1.5\, UCB_t(1 - F(m_2)) < 3.5\, UCB_t(1 - F(m_3))$.*

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

**Algorithm 3** Generic Inner Algorithm B

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

### 2.2.4 Perturbed Linear Bandit

In this subsection, we first introduce a perturbed linear bandit (PLB) framework, then show that our single-episode pricing problem can be formulated as a PLB. Furthermore, the pro-

posed M-LinUCB for PLB is shown to be equivalent to the Inner Algorithm B with a specific UCB construction.

We say the reward $Z_t$, the parameter $\xi_t$, the action set $\mathcal{A}_t$ form a perturbed linear bandit with a perturbation constant $C_p$ if $Z_t = \langle \xi_t, A_t \rangle + \eta_t$ with any selected action $A_t \in \mathcal{A}_t$ and $||\xi_s - \xi_t||_\infty \leq C_p$ for any $s, t$. Here $\eta_t$ is a sub-Gaussian conditional on the filtration $\mathcal{F}_{t-1} = \sigma(\xi_1, A_1, Z_1, \ldots, \xi_t, A_t)$. Note that the condition on the linear parameters $\xi_t$'s implies the existence of a $\xi^*$ such that $||\xi_t - \xi^*||_\infty \leq \frac{C_p}{2}$ for any $t$. Thus the linear parameter $\xi_t$ regulating the reward structure at time $t$ can be viewed as a perturbation from a "central" parameter $\xi^*$. Note that the linear bandit (Abbasi-Yadkori et al., 2011; Chu et al., 2011; Agrawal and Goyal, 2013) is a special zero-perturbation PLB with $\xi_t = \xi^*$ 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 $\xi_t = (1 - F(m_1 + x_t^\top \hat{\theta} - x_t^\top \theta_0), \ldots, 1 - F(m_d + x_t^\top \hat{\theta} - x_t^\top \theta_0))^\top \in \mathbb{R}^d$, which turns out to regulate the reward at time $t$ as shown in Lemma 1 below. Note that for any price $m_j + x_t^\top \hat{\theta} \in \mathcal{S}_t$, the $j$-th element of $\xi_t$ is exactly the purchasing probability of the customer faced with this price. Further define $\xi^* = (1 - F(m_1), \ldots, 1 - F(m_d))^\top$ as the "central" parameter. Then by Lemma 1 below, $\xi_t$'s can be viewed as perturbations from $\xi^*$. It is interesting to see that the perturbations indeed originate from the difference between the estimate $\hat{\theta}$ and the true $\theta_0$, and may change with covariates $x_t$'s.

To transform price setting into action selection, we define a mapping from any price $p = m_j + x_t^\top \hat{\theta} \in \mathcal{S}_t$ to a vector $Q_t(p) \in \mathbb{R}^d$ with $Q_t(p)_j = m_j + x_t^\top \hat{\theta}$ and $Q_t(p)_i = 0, \forall i \neq j$. Namely, $Q_t$ maps a price $p = m_j + x_t^\top \hat{\theta} \in \mathcal{S}_t$ to a vector with a single nonzero $j$-th 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 $A_t = Q_t(p_t)$. Then setting any price $p_t \in \mathcal{S}_t$ means selecting an action $A_t = Q_t(p_t) \in \mathcal{A}_t$ and vice versa. With all these preparations, the following Lemma 1 rigorously forms our single-episode pricing problem into a perturbed linear bandit.

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

**Lemma 1.** *Under Assumption 1, $||\xi_t - \xi^*||_\infty \leq L||\hat{\theta} - \theta_0||_1, \forall t \in [T_0]$. Moreover, under the price-action coupling $A_t = Q_t(p_t)$, the reward $Z_t = p_t 1_{\{v_t \geq p_t\}}$, the parameter $\xi_t$ and the*

*action set $\mathcal{A}_t$ form a perturbed linear bandit with a perturbation constant $2L||\hat{\theta} - \theta_0||_1$.*

Lemma 1 implies that the perturbation is proportional to the $\ell_1$ estimation error $||\hat{\theta}-\theta_0||_1$. If the estimate $\hat{\theta} = \theta_0$, then $\xi_t = \xi^*$ with zero perturbation and the PLB reduces to a classic linear bandit. On the other hand, a worse $\hat{\theta}$ implies a larger perturbation, thus incurring more difficulty in solving the PLB and potentially leading to a larger regret.

According to Lemma 1, $Z_t = A_t^\top \xi_t + \eta_t$ with $||\xi_t - \xi^*||_\infty \le L||\hat{\theta} - \theta_0||_1$, and hence $\xi^*$ can be estimated from historical data. Similar to that in linear bandit (Lattimore and Szepesvári, 2020), we employ the ridge estimator $\hat{\xi}_{t-1} = V_{t-1}(\lambda)^{-1} \sum_{s=1}^{t-1} A_s Z_s$, where $V_{t-1}(\lambda) = \lambda I + \sum_{s=1}^{t-1} A_s A_s^\top$ with the tuning parameter $\lambda > 0$. Denote $H_t = j$ if we pull arm $j$ at time $t$ and $\mathcal{U}_{t-1,j} = \{s : 1 \le s \le t-1, H_s = j\}$. Since $A_s$'s have a single nonzero element and $V_{t-1}(\lambda)$ is a diagonal matrix, we obtain the explicit form for the $j$-th element of $\hat{\xi}_{t-1}$ as $\hat{\xi}_{t-1,j} = \frac{\sum_{s\in\mathcal{U}_{t-1,j}} p_s^2 y_s}{\lambda + \sum_{s\in\mathcal{U}_{t-1,j}} p_s^2}$, which serves as the estimate $\mathrm{EST}_t(1 - F(m_j))$ for $1 - F(m_j) = \xi_j^*$.

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 above estimate $\mathrm{EST}_t(1 - F(m_j)) = \hat{\xi}_{t-1,j}$. The common confidence set $\mathcal{C}_t(\beta_t) = \{\xi \in \mathbb{R}^d : ||\xi - \hat{\xi}_{t-1}||^2_{V_{t-1}(\lambda)} \le \beta_t\}$ yields a marginal confidence radius for each $\hat{\xi}_{t-1,j}$. Due to the simple form of $V_{t-1}(\lambda)$, we obtain an explicit form $\mathrm{CR}_t(1 - F(m_j)) = \sqrt{\frac{\beta_t}{\lambda + \sum_{s\in\mathcal{U}_{t-1,j}} p_s^2}}$. 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, 2020), we specify the parameter $\beta_t = \beta_t^* = p_{\max}^2(1 \vee (\frac{1}{p_{\max}}\sqrt{\lambda d} + \sqrt{2\log(\frac{1}{\delta}) + d\log(\frac{d\lambda + (t-1)p_{\max}^2}{d\lambda})})^2)$. Here $1-\delta$ is the confidence level and $\delta = \frac{1}{T_0}$ is a typical choice (Lattimore and Szepesvári, 2020) with known $T_0$. Thus we use $\delta = \frac{1}{2^{k-2}\ell_2}$ for the application of Inner Algorithm B to the $k$-th episode with an expected 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 sub-algorithms across episodes, one applying classification for linear parameter estimation and the other adapting the UCB idea for online pricing.

Finally, we would like to mention that the proposed perturbed linear bandit framework

**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 $\{x_t\}_{t\in[T]}$

**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 $y_t$.

Apply logistic regression on the data $\{(x_t, p_t, y_t)\}_{t\in[\ell_1]}$ to obtain

$$(\hat{c}_1, \hat{\beta}_1^\top, \hat{b}_1) = \arg\min_{(c,\beta^\top,b)} \sum_{t=1}^{\ell_1} \log(1 + \exp((2y_t - 1)(c, \beta^\top, b)(1, x_t^\top, p_t)^\top)).$$

Construct an estimate $\hat{\theta}_1 = \mathrm{Proj}_\Theta(-\frac{\hat{\beta}_1}{\hat{b}_1})$, where $\Theta = \{\theta \in \mathbb{R}^{d_0} : ||\theta||_1 \le W\}$.

**For** episodes $k = 2, 3, \ldots, n = n(T, \alpha_1, \alpha_2)$, **do**

Cut the interval $G(\hat{\theta}_{k-1}) = [-||\hat{\theta}_{k-1}||_1, p_{\max} + ||\hat{\theta}_{k-1}||_1]$ into $d_k = C\lceil(2^{k-2}\ell_2)^{\frac{1}{6}}\rceil$ same-length intervals with midpoints $m_1, \ldots, m_{d_k}$. Here $\ell_2 = \ell_2(T, \alpha_1, \alpha_2) = \alpha_2$.

**For** $t = t_k + 1, \ldots, t_{k+1}$, where $t_k = \sum_{i=1}^{k-1} \ell_i, t_{k+1} = \sum_{i=1}^{k} \ell_i$ and $\ell_i = \ell_i(T, \alpha_1, \alpha_2)$, **do**

Construct the candidate set $\mathcal{S}_t = \{m_j + x_t^\top \hat{\theta}_{k-1} | j \in [d_k], m_j + x_t^\top \hat{\theta}_{k-1} \in (0, p_{\max})\}$.

Determine the arm set $\mathcal{B}_t = \{j \in [d_k] : \exists p \in \mathcal{S}_t \text{ such that } p = m_j + x_t^\top \hat{\theta}_{k-1}\}$.

Let $\beta_t^* = p_{\max}^2(1 \vee (\frac{1}{p_{\max}}\sqrt{\lambda d_k} + \sqrt{2\log(2^{k-2}\ell_2) + d_k \log(\frac{d_k\lambda + (t-t_k-1)p_{\max}^2}{d_k\lambda})})^2)$.

**For** $j \in \mathcal{B}_t$, **do**

Determine $\mathcal{U}_{t-1,j,k} = \{s : t_k + 1 \le s \le t - 1, H_s = j\}$.

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

Select $j_t \in \arg\max_{j\in\mathcal{B}_t}(m_j + x_t^\top \hat{\theta}_{k-1})\mathrm{UCB}_t(1 - F(m_j))$ and let $H_t = j_t$.

Set the price $p_t = m_{j_t} + x_t^\top \hat{\theta}_{k-1}$ and receive a binary response $y_t$.

Apply logistic regression on the data $\{(x_t, p_t, y_t)\}_{t_k+1\le t\le t_{k+1}}$ to obtain

$$(\hat{c}_k, \hat{\beta}_k^\top, \hat{b}_k) = \arg\min_{(c,\beta^\top,b)} \sum_{t=t_k+1}^{t_{k+1}} \log(1 + \exp((2y_t - 1)(c, \beta^\top, b)(1, x_t^\top, p_t)^\top)).$$

Construct an estimate $\hat{\theta}_k = \mathrm{Proj}_\Theta(-\frac{\hat{\beta}_k}{\hat{b}_k})$.

can be used beyond the above contextual dynamic pricing problem. This motivates us to introduce a general algorithm called M-LinUCB in Algorithm 5 for the perturbed linear bandit framework $Z_t = \langle \xi_t, A_t \rangle + \eta_t$ when any potential action has only one nonzero element. For any vector $v$ with a single nonzero element, denote $\delta(v)$ as the index of this nonzero element. For instance, $\delta((0, 1, 0)^\top) = 2$. Further define $\tilde{\mathcal{B}}_t = \{\delta(a) : a \in \mathcal{A}_t\}$ as the nonzero index set of all potential actions at time $t$ and $\tilde{\mathcal{B}}_t' = \{\delta(A_s) : s \in [t-1]\}$ 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 below.

**Algorithm 5** M-LinUCB for Perturbed Linear Bandit

**Input: (arrives over time)** action sets $\mathcal{A}_t$, $\lambda, \{\beta_t\}_{t\in[T]}$
**For** $t = 1, \ldots, T$, **do**
  Determine $\tilde{\mathcal{B}}_t = \{\delta(a) : a \in \mathcal{A}_t\}$ and $\tilde{\mathcal{B}}'_t = \{\delta(A_s) : s \in [t-1]\}$ where $\delta(v)$ maps any vector with only one nonzero element to its nonzero element index.
  **If** $\tilde{\mathcal{B}}_t \not\subseteq \tilde{\mathcal{B}}'_t$ **do**
    Choose an arbitrary $A_t \in \mathcal{A}_t$ such that $\delta(A_t) \notin \tilde{\mathcal{B}}'_t$.
  **If** $\tilde{\mathcal{B}}_t \subseteq \tilde{\mathcal{B}}'_t$ **do**
    Calculate $\text{LinUCB}_t(a) = \max_{\xi\in\mathcal{C}_t(\beta_t)}\langle\xi, a\rangle$ for $a \in \mathcal{A}_t$ where $\mathcal{C}_t(\beta_t) = \{\xi \in \mathbb{R}^d : ||\xi - \hat{\xi}_{t-1}||^2_{V_{t-1}(\lambda)} \le \beta_t\}$ and $\hat{\xi}_{t-1} = V_{t-1}(\lambda)^{-1}\sum_{s=1}^{t-1} A_s Z_s, V_{t-1}(\lambda) = \lambda I + \sum_{s=1}^{t-1} A_s A_s^\top$.
    Choose $A_t \in \arg\max_{a\in\mathcal{A}_t} \text{LinUCB}_t(a)$.
  Receive a reward $Z_t$.

**Lemma 2.** *Applying Algorithm 5 to the PLB formulation of our single-episode pricing problem with* $\beta_t = \beta_t^* = p_{\max}^2(1 \vee (\frac{1}{p_{\max}}\sqrt{\lambda d} + \sqrt{2\log(\frac{1}{\delta}) + d\log(\frac{d\lambda+(t-1)p_{\max}^2}{d\lambda})})^2)$ *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 of the proposed DIP policy. As DIP divides the total time horizon into episodes, we first conduct the regret analysis on a single episode and then merge them together. For the single episode, our discretization procedure leads to a natural decomposition of the regret into a discrete part and a continuous part. 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 framework. Then we proceed to analyze the continuous-part regret and form the overall regret bound of DIP.

In our single-episode regret analysis, we denote the total horizon as $[T_0]$ and use $\hat{\theta}$ 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 $\mathcal{S}_t$, i.e., $\tilde{p}_t^* \in \arg\max_{p\in\mathcal{S}_t} p(1-F(p-x_t^\top\theta_0))$. Thus the regret $r_t$ in (1) can be rewritten as

$$\underbrace{\tilde{p}_t^*(1-F(\tilde{p}_t^*-x_t^\top\theta_0))-p_t(1-F(p_t-x_t^\top\theta_0))}_{r_{t,1}}+\underbrace{p_t^*(1-F(p_t^*-x_t^\top\theta_0))-\tilde{p}_t^*(1-F(\tilde{p}_t^*-x_t^\top\theta_0))}_{r_{t,2}}.$$

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-part 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.

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 $A_t^*=Q_t(\tilde{p}_t^*)$. Therefore, the selected action $A_t=Q_t(p_t)$ yields the regret $\tilde{p}_t^*(1-F(\tilde{p}_t^*-x_t^\top\theta_0))-p_t(1-F(p_t-x_t^\top\theta_0))$ 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 consider a PLB setting with the reward model $Z_t=\langle\xi_t,A_t\rangle+\eta_t$ which satisfies the following conditions.

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

**Condition 2** For any $t\in\mathbb{N}^+$, $||\xi_t||_\infty\le C_1$.

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

**Condition 4** For any $t\in\mathbb{N}^+$, $\eta_t$ is a 1-conditionally sub-Gaussian random variable, i.e., $\mathbb{E}(\exp(\alpha\eta_t)|\mathcal{F}_{t-1})\le\exp(\frac{\alpha^2}{2})$, where $\mathcal{F}_{t-1}=\sigma(\xi_1,A_1,Z_1,\ldots,\xi_t,A_t)$.

**Remark 1.** *Condition 1 ensures a constant regret upper bound at each time and is commonly adopted in linear bandit (Lattimore and Szepesvári, 2020). Condition 2 assumes the infinity norm of $\xi_t$ to be bounded. Condition 3 implies there is only one nonzero element bounded in absolute value for any action. Condition 4 implies that the noise is sub-Gaussian conditional on all the past parameters, actions, rewards, as well as the current parameter and action. As we will show later, the perturbed linear bandit formulation of our single-episode pricing problem satisfy 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 $\beta_t = \tilde{\beta}_t = 1 \vee (C_1\sqrt{\lambda d} + \sqrt{2\log(\frac{1}{\delta}) + d\log(\frac{d\lambda+(t-1)a_{\max}^2}{d\lambda})})^2$ has the regret bound*

$$R_{T_0}^{PLB} \leq 2\sqrt{2dT_0\tilde{\beta}_{T_0}\log(\frac{d\lambda + T_0 a_{\max}^2}{d\lambda})} + 2a_{\max}C_pT_0 + 2d.$$

**Proof Sketch**: We construct a new sequence $\{\dot{\xi}_t\}_{2\leq t\leq T_0}$ and control the "pseudo-regret" $\sum_{t=2}^{T_0}\langle\dot{\xi}_t, \dot{A}_t - A_t\rangle$ with sub-linear order $\tilde{\mathcal{O}}(\sqrt{T_0})$, where $\dot{A}_t = \arg\max_{a\in\mathcal{A}_t}\langle\dot{\xi}_t, a\rangle$. By closeness of $\dot{\xi}_t$ and $\xi_t$ 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 $\{\dot{\xi}_t\}_{2\leq t\leq T_0}$ and rigorous proofs are deferred to Section A of the Supplement. ■

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. On the other hand, when $C_p$ is too large, the perturbed linear bandit would become intractable. Thus the cardinality of $C_p$ plays an important role in the overall regret. Interestingly, by Lemma 1, this perturbation constant in the PLB formulation of our single-episode pricing problem is proportional to $||\hat{\theta} - \theta_0||_1$, which matches with the intuition that a larger estimation error of the input $\hat{\theta}$ would lead to more revenue loss in this episode. Moreover, Lemma 3 might be of independent interest since it provides an informative decomposition of the regret bound for a general PLB problem.

**Remark 2.** *Our proposed PLB can be viewed as a misspecified linear bandit (Lattimore et al., 2020; Pacchiano et al., 2020; Foster et al., 2020) with a misspecification level $\epsilon_* = a_{\max}C_p/2$, where the latter has a general regret of $\tilde{\mathcal{O}}(d\sqrt{T_0} + \epsilon_*\sqrt{d}T_0)$. On the other hand, by utilizing the special structure of our PLB setting, we prove a regret of $\tilde{\mathcal{O}}(d\sqrt{T_0} + a_{\max}C_pT_0) = \tilde{\mathcal{O}}(d\sqrt{T_0} + \epsilon_*T_0)$, which avoids the dependence of $d$ on the linear term. Such $\sqrt{d}$ improvement is critical as $d$ refers to the discretization cardinality in our problem and $d = C'\lceil T_0^{1/6}\rceil$ is an optimal choice for the tradeoff between discrete and continuous parts of the regret. Moreover, we achieve better regret by applying M-LinUCB, which is much simpler than the algorithms for misspecified linear bandits.*

**Remark 3.** *Non-stationary linear bandits (NLB) (Cheung et al., 2018; Russac et al., 2019; Zhao et al., 2020) also allow changing linear parameters $\xi_t$ but design policies to adapt to the smooth variations $B_{T_0} = \sum_{t=1}^{T_0-1} ||\xi_t - \xi_{t+1}||_2$. Our PLB setting fits an NLB with linear variations $B_{T_0} = \mathcal{O}(C_p T_0)$. The nonasymptotic results in Cheung et al. (2018); Zhao et al. (2020) 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 proved Lemma 3 provides regret behaviors with a fixed $T_0$ for $C_p \to 0$, i.e., 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. (2018); Zhao et al. (2020) also yield regrets for fixed $T_0$ and $C_p \to 0$, they suggest worse regrets such as $\tilde{\mathcal{O}}(wC_pT_0 + \frac{T_0}{\sqrt{w}})$ ($w$ 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 will inevitably deteriorate the performance guarantee for our pricing problem.*

Next we prove an $\Omega(C_pT_0)$ regret lower bound for the PLB with a perturbation $C_p$. This implies that the linear term in the upper bound is inevitable due to the potentially adversarial perturbations. Define $PB(\tilde{\xi}, C_p) = \{\xi \in \mathbb{R}^d : ||\xi - \tilde{\xi}||_\infty \leq \frac{C_p}{2}\}$ as a parameter set with respect to a "central" parameter $\tilde{\xi}$ and a perturbation quantification $C_p$.

**Proposition 1.** *For any PLB algorithm $\mathcal{A}^*$, any $\tilde{\xi}$ with all positive elements and $\frac{C_p}{2} < \min_{i\in[d]} \tilde{\xi}_i$, there exists a PLB with parameters $(\xi_1, \dots, \xi_t, \dots)$ and action sets $(\mathcal{A}_1, \dots, \mathcal{A}_t, \dots)$ satisfying $\xi_t \in PB(\tilde{\xi}, C_p), \forall t \in \mathbb{N}^+$ and a constant $C_0$ only dependent on $\tilde{\xi}$ such that*

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

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 $\frac{1}{p_{\max}}$ as $\tilde{\xi}_t = \frac{1}{p_{\max}}\xi_t$, $\tilde{Z}_t = \frac{1}{p_{\max}} Z_t$, $\tilde{\eta}_t = \frac{1}{p_{\max}}\eta_t$, we obtain the transformed model $\tilde{Z}_t = \langle \tilde{\xi}_t, A_t \rangle + \tilde{\eta}_t$ with the perturbation constant $\tilde{C}_p = \frac{2L||\hat{\theta} - \theta_0||_1}{p_{\max}}$, which satisfies Conditions 1-4. 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 $\beta_t = \tilde{\beta}_t$ on the transformed model, with their regrets admitting a scaling relationship. By formalizing the above reasoning, we obtain the following Theorem 1.

**Theorem 1.** *Suppose Assumption 1 holds. 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*

$$R_{T_0,1} \leq 2\sqrt{2dT_0\beta^*_{T_0}\log(\frac{d\lambda+T_0p^2_{\max}}{d\lambda})} + 4p_{\max}L||\hat{\theta}-\theta_0||_1T_0 + 2dp_{\max}.$$

Theorem 1 provides an upper bound of the discrete-part regret on a single episode. The first term is sub-linear as $\tilde{\mathcal{O}}(\sqrt{T_0})$ while the second term is linear in $T_0$ and proportional to the estimation error $||\hat{\theta}-\theta_0||_1$, which invokes the perturbation in our PLB formulation. The third term will be dominated by the first two terms as we further specify $d$ to yield a best tradeoff between discrete and continuous parts of the regret.

It remains to bound the continuous-part regret which is the expected reward difference between the discrete best prices $\tilde{p}^*_t$ and overall best prices $p^*_t$, i.e., $r_{t,2} = p^*_t(1-F(p^*_t - x_t^\top\theta_0)) - \tilde{p}^*_t(1-F(\tilde{p}^*_t - x_t^\top\theta_0))$. Denote $f_q(p) = p(1-F(p-q))$. Then $r_{t,2}$ can be rewritten as $f_{x_t^\top\theta_0}(p^*_t) - f_{x_t^\top\theta_0}(\tilde{p}^*_t)$. We adopt the following Assumption 2.

**Assumption 2.** *There exists a constant* $C$ *such that for any* $q = x^\top\theta_0$ *and* $x \in \mathcal{X}$*, we have* $f_q(p^*(x)) - f_q(p) \leq C(p^*(x)-p)^2, \forall p \in [0, p_{\max}]$.

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 $f_{x^\top\theta_0}(p)$ over the local neighborhoods of the maximizers $p^*(x)$. In Proposition 2, by applying the 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 even for locally erratic $F$'s as long as Assumption 2 is satisfied.

**Proposition 2.** *Assumption 2 holds if* $F''(\cdot)$ *is bounded on* $[-||\theta_0||_1, p_{\max}+||\theta_0||_1]$.

We now discuss how to derive a bound for the continuous-part regret under Assumption 2. By our discretization approach, $\{m_i + x_t^\top\hat{\theta}\}_{i\in[d]}$ are a sequence of points that "cover" $[0, p_{\max}]$ with equal adjacent distance $\frac{p_{\max}+2\|\hat{\theta}\|_1}{d}$. Since $\mathcal{S}_t = \{m_j + x_t^\top\hat{\theta} | j \in [d], m_j + x_t^\top\hat{\theta} \in (0, p_{\max})\}$ 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 $\mathcal{O}(\frac{1}{d^2})$ according to Assumption 2. Since 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 $r_{t,2} = \mathcal{O}(\frac{1}{d^2})$. This leads to the total regret in a single-episode pricing problem.

**Theorem 2.** *Suppose Assumptions 1 – 2 hold. With probability at least $1-\delta$, applying Algorithm 3 on the single-episode pricing problem yields the total regret $R_{T_0}$ satisfying*

$$R_{T_0} \leq 2\sqrt{2dT_0\beta_{T_0}^* \log(\frac{d\lambda + T_0 p_{\max}^2}{d\lambda})} + 4p_{\max}L||\hat{\theta} - \theta_0||_1 T_0 + C\frac{T_0}{d^2} + 2dp_{\max},$$

*where $C$ is a constant dependent on $W$. Moreover, by setting $\delta = \frac{1}{T_0}, d = C'\lceil T_0^{1/6} \rceil$ and taking the expectation, we have $\mathbb{E}(R_{T_0}) = \tilde{\mathcal{O}}(T_0^{2/3}) + 4p_{\max}L||\hat{\theta} - \theta_0||_1 T_0$.*

In Theorem 2, we prove a high probability bound as well as an expected regret bound for our pricing policy on a single episode. The expected regret is bounded by a sub-linear $\tilde{O}(T^{2/3})$ term and a linear term proportional to the $\ell_1$ estimation error $||\hat{\theta} - \theta_0||_1$. Our DIP policy applies Algorithm 3 to each $k$-th episode with $\hat{\theta} = \hat{\theta}_{k-1}$, for $k = 2, \ldots, n$. Thus by applying Theorem 2 to each episode, we obtain the final regret bound over the whole horizon.

**Theorem 3.** *Suppose Assumptions 1 – 2 hold. By choosing $\alpha_1, \alpha_2, C, \lambda$ as some constants, our DIP policy as in Algorithm 4 yields the expected regret*

$$\mathbb{E}(R_T) = \tilde{\mathcal{O}}(T^{2/3}) + 4p_{\max}L\sum_{k=2}^{n} 2^{k-2}\ell_2 \mathbb{E}||\hat{\theta}_{k-1} - \theta_0||_1.$$

In Theorem 3, the first part is $\tilde{\mathcal{O}}(T^{2/3})$ while the second part depends on the sequence of estimation errors $||\hat{\theta}_k - \theta_0||_1$. If the estimates $\{\hat{\theta}_k\}_{k\in[n-1]}$ are perfectly accurate, the second term vanishes and the overall regret is of $\tilde{\mathcal{O}}(T^{2/3})$. According to the non-asymptotic error bound of lasso-penalized logistic regression (Hastie et al., 2015; Wainwright, 2019), with additional conditions such as data regularity, it is possible to achieve $||\hat{\theta}_k - \theta_0||_1 = \mathcal{O}_p(\ell_k^{-1/2})$. This yields an $\tilde{\mathcal{O}}(T^{1/2})$ second term and results in an overall regret of $\tilde{\mathcal{O}}(T^{2/3})$. In general, if $\mathbb{E}||\hat{\theta}_k - \theta_0||_1 = \mathcal{O}(\ell_k^{-\alpha})$ for some $0 < \alpha \leq \frac{1}{2}$, then by the doubling construction, we can conclude $\sum_{k=2}^{n} 2^{k-2}\ell_2 \mathbb{E}||\hat{\theta}_{k-1} - \theta_0||_1 = \mathcal{O}(T^{1-\alpha})$. Thus the overall regret is of $\tilde{\mathcal{O}}(T^{\frac{2}{3}\vee(1-\alpha)})$. Therefore, Theorem 3 presents how DIP absorbs estimation errors into its regret when the noise distribution is unknown. In all of our experiments, the logistic regression employed for

$\theta_0$ estimation is very accurate. Thus, the overall regret mainly emerges from the M-LinUCB driving the exploration and exploitation of the unknown $F$.

**Remark 4.** *At first glance, the obtained regret upper bound is worse than the typical $\mathcal{O}(T^{1/2})$ lower bound in linear bandit (Lattimore and Szepesvári, 2020) and dynamic pricing with known noise distribution (Javanmard and Nazerzadeh, 2019). However, we would like to point out that our problem involves both unknown linear parameter $\theta_0$ 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 (2021) considered a nonparametric pricing problem and proved an $\mathcal{O}(T^{(d_0+2)/(d_0+4)})$ lower bound under some additional smoothness and concavity assumptions. Here $d_0$ is the dimension of the nonparametric component. In our pricing problem with a one-dimensional nonparametric component $F$ and an additional unknown $\theta_0$, the lower bound should be at least $\mathcal{O}(T^{3/5})$, which is larger than the typical $\mathcal{O}(T^{1/2})$ lower bound. We leave the investigation of lower bound of our problem for future work.*

# 4 Simulation Study

We demonstrate the performance of our DIP policy on synthetic datasets and compare it with RMLP and RMLP-2 proposed by Javanmard and Nazerzadeh (2019). The implementation details of DIP, RMLP and RMLP-2 are provided in Section B of the Supplement.

Let $\Phi(\mu, \sigma^2)$ denote the CDF of $N(\mu, \sigma^2)$ distribution. For the first six examples, we consider a scalar covariate $x_t \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, 1]$ and set $\theta_0 = 30$. The CDF $F$ of the noise distribution is designed as follows, where Examples 1 and 5 are motivated from the real application in Section 5. Their probability density functions (PDF) are shown in Figure 4.

**Example 1.** The true $F = \frac{1}{2}\Phi(-4, 6) + \frac{1}{2}\Phi(4, 6)$.

**Example 2.** The true $F = \frac{1}{3}\Phi(-6, \frac{\pi^2}{3}) + \frac{1}{3}\Phi(-1, \frac{\pi^2}{3}) + \frac{1}{6}\Phi(1, \frac{\pi^2}{3}) + \frac{1}{6}\Phi(6, \frac{\pi^2}{3})$.

**Example 3.** The true $F = \frac{1}{4}\Phi(-7, \frac{\pi^2}{3}) + \frac{1}{4}\Phi(-3, \frac{\pi^2}{3}) + \frac{1}{4}\Phi(3, \frac{\pi^2}{3}) + \frac{1}{4}\Phi(7, \frac{\pi^2}{3})$.

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

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

**Example 6.** The true $F = \frac{1}{2}\Phi(-2.5, 5) + \frac{1}{2}\Phi(2.5, 5)$.

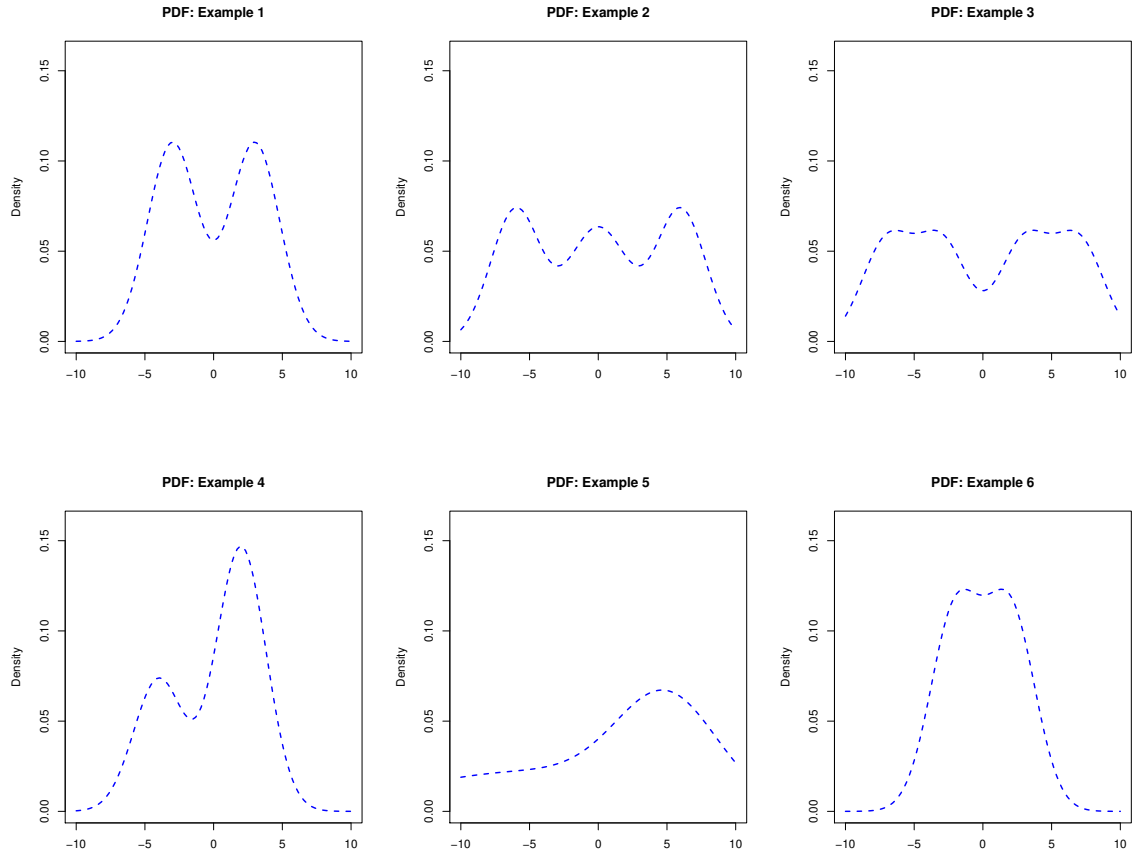


Figure 4: PDFs of the noise distribution in Examples 1-6.

As shown in Figure 4, Examples 1-3 have symmetric PDFs with two, three and four modes, respectively, while Examples 4-5 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. Due to 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, where the PDF is asymmetric and has heavy tails. On the other hand, our DIP policy gradually learns $F$ in the pricing process and achieves sub-linear 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.

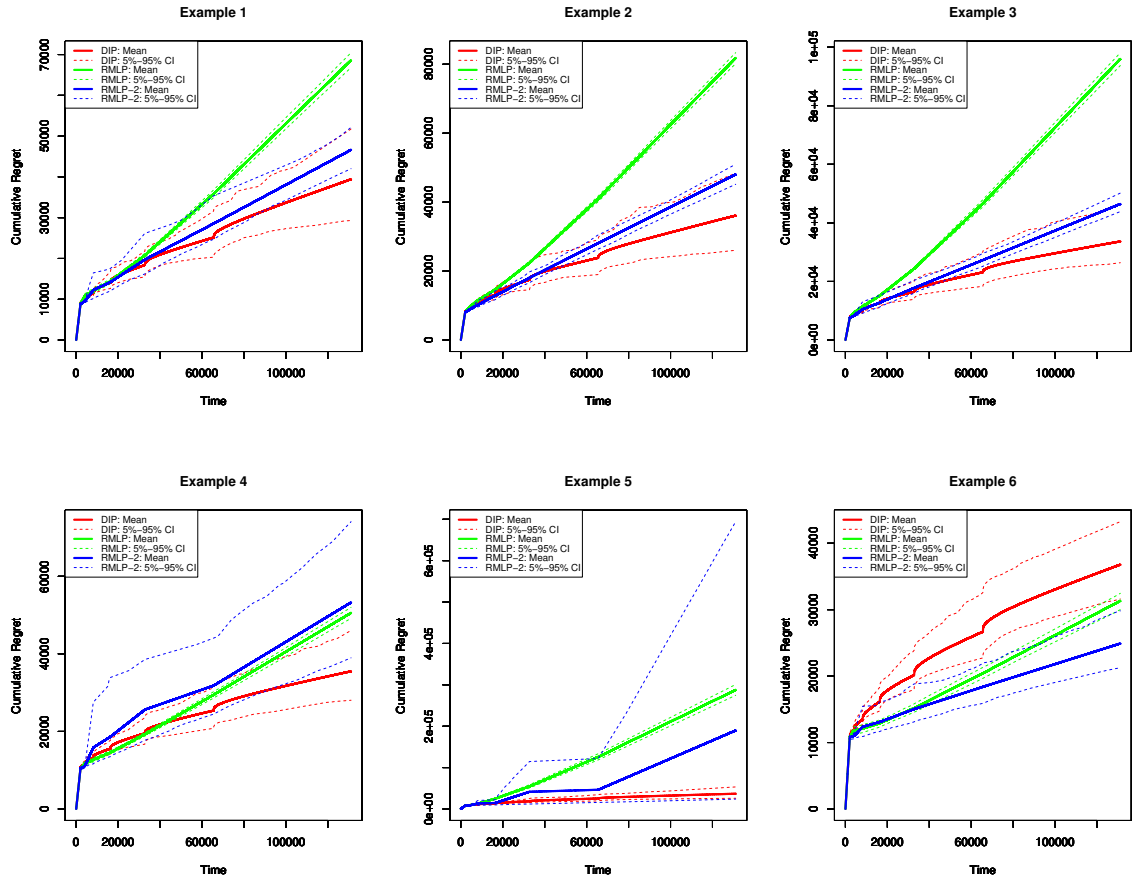


Figure 5: Regret comparisons of DIP, RMLP and RMLP-2 in Examples 1-6.

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 $\theta_0$, and the context generation distribution.

**Example 7.** Dimension $d_0 = 3$, $\theta_0 = (10, 10, 10)^\top$, $x_t \overset{\text{i.i.d.}}{\sim} \text{Unif}[0.3, 1]^3$.

**Example 8.** Dimension $d_0 = 10$, $\theta_0 = (3, \ldots, 3)^\top$, $x_t \overset{\text{i.i.d.}}{\sim} \text{Unif}[0.1, 1]^{10}$.

**Example 9.** Dimension $d_0 = 10$, $\theta_0 = (3, \ldots, 3)^\top$, $x_t \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, 1]^{10}$.

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 due to the unsatisfactory performance of RMLP. For all three methods, we estimate $\theta_0$ in each of six episodes and use it for pricing in subsequent episode. Figure 7 shows boxplots of estimation errors $\|\hat{\theta}_k - \theta_0\|_2$ for all six episodes $k = 1, \ldots, 6$ and all three methods in Examples 7-9. In Examples 7-8, 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 since it specifies $F$ as $\frac{\exp(x)}{\exp(x)+1}$ with the variance $\frac{\pi^2}{3}$, 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 over 80 in episode 2. Then in this replication, the regret of RMLP-2 in the subsequent episode 3 can be large due to this unsatisfactory estimate. We conjecture that the unstable estimations of RMLP-2 in Examples 7-8 are due to its produced singular price-covariate data in each episode. In Section C of the Supplement, 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.

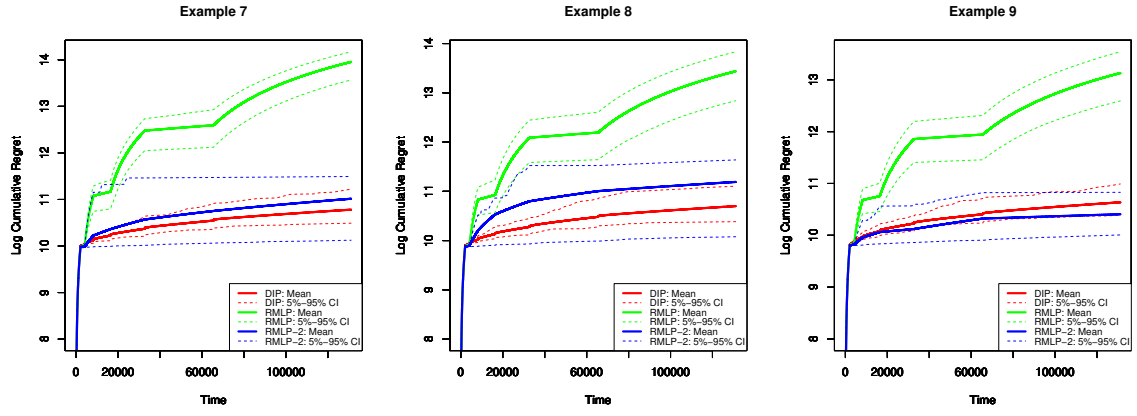


Figure 6: Log regret comparisons of DIP, RMLP and RMLP-2 in Examples 7-9.

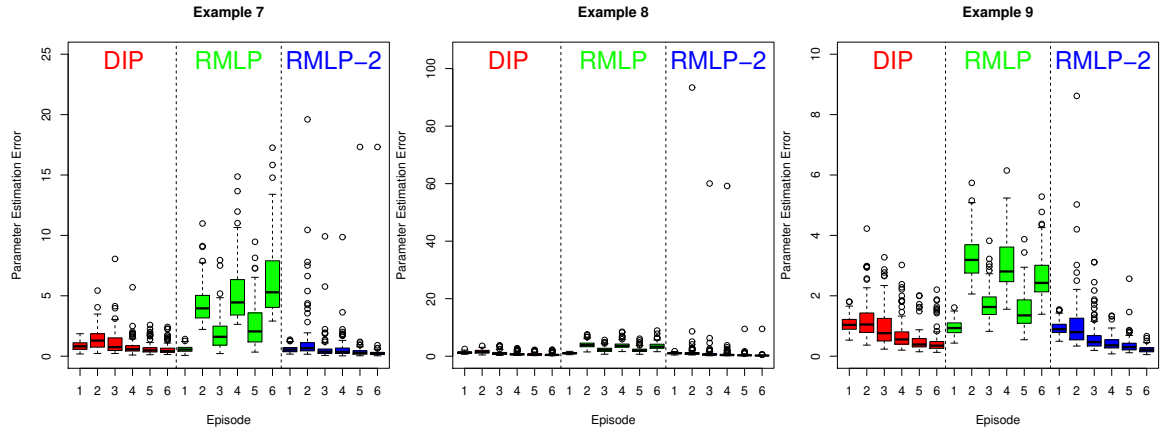


Figure 7: Estimation errors $\|\hat{\theta}_k - \theta_0\|_2$ of DIP, RMLP and RMLP-2 over six episodes in Examples 7-9.

# 5 Real Data Analysis

We explore the efficiency of our proposed DIP policy on a real-life auto loan dataset provided by the Center for Pricing and Revenue Management at Columbia University. This dataset was first studied by Phillips et al. (2015) and further used by Bastani et al. (2019) and Ban and Keskin (2020) to evaluate different dynamic pricing algorithms.

The dataset 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 or not the price was accepted by the borrower, i.e., the customer's binary purchasing decision in our model.

We adopt the feature selection result used in Bastani et al. (2019) and Ban and Keskin (2020) 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, i.e., $p = \text{Monthly Payment} \times \sum_{\tau=1}^{\text{Term}} (1 + \text{Rate})^{-\tau} - \text{Loan Amount}$.

We use one thousand dollars 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 while data were collected. Thus we follow the off-policy learning idea used in Bastani et al. (2019); Ban and Keskin (2020) to first estimate the customer choice model using the entire dataset and use it as the grand truth to generate the willingness-to-pay of each customer given any prices. We utilize a two-step estimation procedure to estimate the unknown $\theta_0$ and $F$. In particular, we use logistic regression to estimate $\theta_0$ and then use the kernel density estimation idea to estimate $F$. The details of this estimation procedure are deferred to Section D of the Supplement. The estimated noise PDF for the US is shown in the left plot of Figure 8. The estimated $\hat{\theta}_0$ and $\hat{F}$ are treated as the true parameters for the customer choice model $y_t \sim \text{Ber}(1 - \hat{F}(p_t - x_t^\top \hat{\theta}_0))$. Note that these true 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.

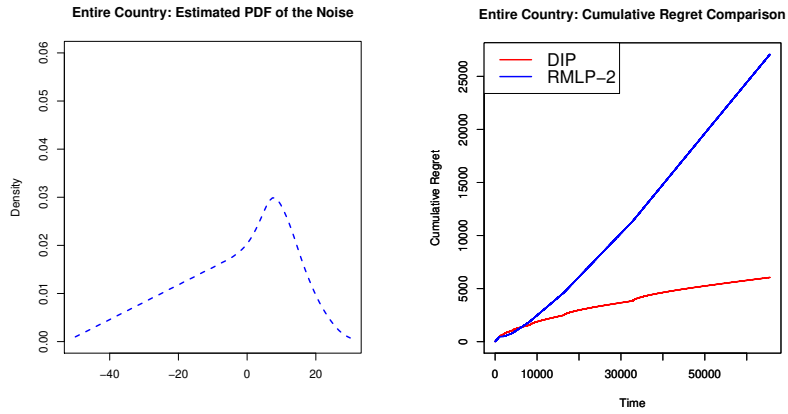


Figure 8: The left plots shows the PDF of the noise distribution for whole US data and the right plot shows the regret comparison of DIP and RMLP-2.

We only compare DIP with RMLP-2 since RMLP-2 is more robust than RMLP as shown in synthetic data in Section 4. Since the dimension is low and the coefficients are nonsparse, we apply the RMLP-2 policy 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 208085 for 50 times and apply DIP and RMLP-2 policy to each

of the 50 replications and then record the average cumulative regrets.

As shown in the right plot of Figure 8, DIP 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 sub-linear cumulative regret while 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 $\theta_0$ estimation since it invests a certain amount in price explorations and generates a more well-distributed dataset. The 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 leading to unsatisfactory estimates. This phenomenon is similar to that shown in Examples 7-8 of the synthetic experiments.

Next we evaluate the performance of DIP and RMLP-2 by focusing on data in California which has nearly 30000 applications. We apply the same estimation procedure for $\theta_0$ and $F$ on the California dataset to obtain the true customer choice model for California. The estimated PDF of the noise distribution for the California data is shown in the left panel of Figure 9. It has a multimodal pattern and does not satisfy the log-concave condition required by RMLP-2. 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 9, DIP again achieves a sub-linear regret, which outperforms that of RMLP-2 eventually.

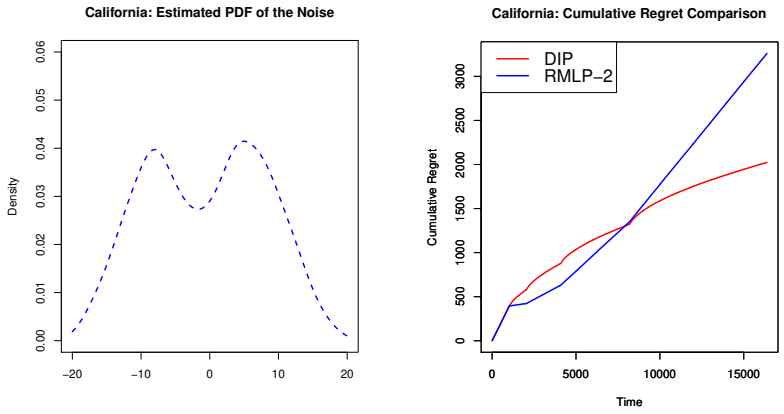


Figure 9: The left plots shows the PDF of the noise distribution for the California data and the right plot shows the regret comparison of DIP and RMLP-2.

# 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 sub-linear regret bound. We apply DIP on various synthetic datasets and a real online Auto Lending dataset 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 non-sparse 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., 1993) 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, 2020) into our dynamic pricing regime.

# References


Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning* **47,** 235–256.

Ban, G.-Y. and Keskin, B. (2020). Personalized dynamic pricing with machine learning: High dimensional features and heterogeneous elasticity. *Forthcoming, Management Science* .

Bastani, H., Simchi-Levi, D., and Zhu, R. (2019). Meta dynamic pricing: Transfer learning across experiments. *Available at SSRN 3334629* .

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

Besbes, O. and Zeevi, A. (2011). On the minimax complexity of pricing in a changing environment. *Operations research* **59,** 66–79.

Besbes, O. and Zeevi, A. (2015). On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science* **61,** 723–739.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore.

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

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721* .

Cesa-Bianchi, N., Cesari, T., and Perchet, V. (2019). Dynamic pricing with finitely many unknown valuations. In *Algorithmic Learning Theory*, pages 247–273. PMLR.

Chen, N. and Gallego, G. (2018). A primal-dual learning algorithm for personalized dynamic pricing with an inventory constraint. *Available at SSRN 3301153* .

Chen, N. and Gallego, G. (2021). Nonparametric pricing analytics with customer covariates. *Forthcoming, Operations Research* .

Chen, X., Owen, Z., Pixton, C., and Simchi-Levi, D. (2021). A statistical learning approach to personalization in revenue management. *Forthcoming, Management Science* .

Chen, Y., Wen, Z., and Xie, Y. (2019). Dynamic pricing in an evolving and unknown marketplace. *Available at SSRN 3382957* .

Cheung, W. C., Simchi-Levi, D., and Wang, H. (2017). Dynamic pricing and demand learning with limited price experimentation. *Operations Research* **65,** 1722–1731.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. (2018). Hedging the drift: Learning to optimize under non-stationarity. *Available at SSRN 3261050* .

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214.

Cohen, M. C., Lobel, I., and Paes Leme, R. (2020). Feature-based dynamic pricing. *Management Science* **66,** 4921–4943.

den Boer, A. V. (2015a). Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science* **20,** 1–18.

den Boer, A. V. (2015b). Tracking the market: Dynamic pricing and learning in a changing environment. *European journal of operational research* **247,** 914–927.

den Boer, A. V. and Keskin, N. B. (2020). Discontinuous demand functions: estimation and pricing. *Management Science* **66,** 4516–4534.

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

Foster, D. and Rakhlin, A. (2020). Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR.

Foster, D. J., Gentile, C., Mohri, M., and Zimmert, J. (2020). Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems* **33,**.

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

Golrezaei, N., Javanmard, A., and Mirrokni, V. (2021). Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Operations Research* **69,** 297–314.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Huang, J., Mani, A., and Wang, Z. (2021). The value of price discrimination in large social networks. *Management Science, Forthcoming* .

Javanmard, A. (2017). Perishability of data: dynamic pricing under varying-coefficient models. *The Journal of Machine Learning Research* **18,** 1714–1744.

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

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

Kallus, N. and Zhou, A. (2020). Fairness, welfare, and equity in personalized pricing. *arXiv preprint arXiv:2012.11066* .

Keskin, N. B. and Zeevi, A. (2014). Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research* **62,** 1142–1167.

Keskin, N. B. and Zeevi, A. (2017). Chasing demand: Learning and earning in a changing environment. *Mathematics of Operations Research* **42,** 277–307.

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

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lattimore, T., Szepesvari, C., and Weisz, G. (2020). Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR.

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

Misra, K., Schwartz, E. M., and Abernethy, J. (2019). Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science* **38,** 226–252.

Mueller, J., Syrgkanis, V., and Taddy, M. (2018). Low-rank bandit methods for high-dimensional dynamic pricing. *arXiv preprint arXiv:1801.10242* .

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

Pacchiano, A., Phan, M., Abbasi Yadkori, Y., Rao, A., Zimmert, J., Lattimore, T., and Szepesvari, C. (2020). Model selection in contextual stochastic bandit problems. *Advances in Neural Information Processing Systems* **33,**.

Perchet, V., Rigollet, P., et al. (2013). The multi-armed bandit problem with covariates. *The Annals of Statistics* **41,** 693–721.

Phillips, R., Şimşek, A. S., and Van Ryzin, G. (2015). The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Science* **61,** 1741–1759.

Qiang, S. and Bayati, M. (2016). Dynamic pricing with demand covariates. *Available at SSRN 2765257* .

Russac, Y., Vernade, C., and Cappé, O. (2019). Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, pages 12040–12049.

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

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, J., Shen, X., and Liu, Y. (2008). Probability estimation for large-margin classifiers. *Biometrika* **95,** 149–167.

Wang, Y., Chen, B., and Simchi-Levi, D. (2021). Multimodal dynamic pricing. *Forthcoming, Management Science* .

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

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

Weed, J., Perchet, V., and Rigollet, P. (2016). Online learning in repeated auctions. In *Conference on Learning Theory*, pages 1562–1583. PMLR.

Zhao, P., Zhang, L., Jiang, Y., and Zhou, Z.-H. (2020). A simple approach for non-stationary linear bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 2020.