Online Regularization towards Always-Valid High-Dimensional Dynamic Pricing

Chi-Hua Wang [*], Zhanyu Wang[†], Will Wei Sun[‡], and Guang Cheng[§]

Abstract

Devising a dynamic pricing policy with always valid online statistical learning procedures is an important and as yet unresolved problem. Most existing dynamic pricing policies, which focus on the faithfulness of adopted customer choice models, exhibit a limited capability for adapting to the online uncertainty of learned statistical models during the pricing process. In this paper, we propose a novel approach for designing a dynamic pricing policy based on regularized online statistical learning with theoretical guarantees. The new approach overcomes the challenge of continuous monitoring of the online Lasso procedure and possesses several appealing properties. In particular, we make the decisive observation that the always-validity of pricing decisions builds and thrives on the *online regularization* scheme. Our proposed online regularization scheme equips the proposed optimistic online regularized maximum likelihood pricing (`OORMLP`) pricing policy with three major advantages: encode market noise knowledge into pricing process optimism; empower online statistical learning with always-validity overall decision points; envelop prediction error process with time-uniform non-asymptotic oracle inequalities. This type of non-asymptotic inference results allows us to design more sample-efficient and robust dynamic pricing algorithms in practice. In theory, the proposed `OORMLP` algorithm exploits the sparsity structure of high-dimensional models and secures a logarithmic regret in a decision horizon. These theoretical advances are made possible by proposing an optimistic online Lasso procedure that resolves dynamic pricing problems at the *process* level, based on a novel use of non-asymptotic martingale concentration. In experiments, we evaluate `OORMLP` in different synthetic and real pricing problem settings and demonstrate that `OORMLP` advances the state-of-the-art methods.

Key Words: bandit, dynamic pricing, martingale concentration, online lasso, time-uniform oracle inequality, regret analysis.

[*]Department of Statistics, UCLA, CA, 90095. Email: tsubasa3002101@gmail.com.
[†]Department of Statistics, Purdue University, IN 47907. Email: wang4094@purdue.edu
[‡]Daniels School of Business, Purdue University, IN 47907. Email: sun244@purdue.edu.
[§]Department of Statistics, UCLA, CA, 90095. Email:guangcheng@ucla.edu

1 Introduction

With the growing availability and differentiation of digital products, modern online marketplaces present a unique challenge for dynamic pricing algorithms: they must customize pricing decisions for a diverse range of digital goods to the seller's customer database in an online environment. In response to such a unique challenge, *online* training in modern dynamic pricing systems has increasingly included market knowledge and business insights, such as product features, marketing environment, and customer purchasing behavior. Indeed, dynamic pricing has been employed in a variety of services and businesses, including hospitality, tourism, entertainment, retail, energy, and public transportation (den Boer, 2015), and has evolved into an integral part of revenue management in modern online service industries.

A significant challenge of dynamic pricing in the modern digital economy is making customized pricing decisions for products, services, and solutions on the basis of item-level data. Besides, while most practical scenarios involve high-dimensional item-level data, only a small number of the observed features are typically decisive in the pricing decision process. In addition, high-dimensional dynamic pricing procedure has another layer of complexity: the entire pricing decision-making process is trained and learned from binary feedback. That is, pricing decision makers only observe and learn from the sale status for the price that was delivered, rather than learning from the true market value of the current item. To generate business insights on pricing mechanism, it is desirable to learn models that attributes to small number of decisive pricing factors to enhance explainablity of online learned market value model of products while maximizing the revenue.

Further, risk control of the online learned model on *continuously monitor* dynamic pricing procedure is in emerging demand from industrial practice because the opportunity cost of lengthy pricing experiments is high and regrettable (Johari et al. (2021)). Indeed, it is desirable to detect the true product market value as quickly as possible or to abolish the running pricing experiment if the revenue improvement appears unpromising so that the scientist may

test other available actions. Besides, optimizing the running time in advance is unfortunately impractical due to lacking knowledge on seeking revenue improvement and cost elasticity. In modern dynamic pricing practice, deployment of online statistical learning methodology turns out to be impeded by the such dynamic trade-off between maximum revenue improvement detection and minimum running time. Resolving such trade-off is a crucial advancement in statistical methodology for real-time data and persuades our investigation on the problem of *continuous monitoring high-dimensional dynamic pricing problems*.

Continuous monitoring of high-dimensional dynamic pricing problem is a setting in which decision-makers seek to recover a sparse product market value model and maximize collected revenue (high-dimensional dynamic pricing), while the decision-makers are allowed to terminate the pricing algorithm whenever they wish, *and* the result still maintains statistical validity (continuous monitoring). Such a setting arises naturally in industrial practice (Johari et al., 2021) but remains challenging in the literature, preventing practitioners from effectively deploying high-dimensional statistical methodology effectively in modern online service industries. Specifically, we consider a company that sells products to customers over a *randomly stopped* time horizon. Each period, a new product is introduced, and the dynamic pricing algorithm is responsible for deciding its price. The pricing decision is based on the product feature and the historical pricing and sales data. Once the price is decided, the market either accepts or rejects the product, depending on whether the price is less than or more than the product's market value. The company has no idea what the market value of each product is, other than that it is a function in terms of the value of the product feature (Broder and Rusmevichientong, 2012; Keskin and Zeevi, 2014; Javanmard and Nazerzadeh, 2019). Accordingly, the seller can utilize historical prices and sales data to infer market values for various product features and use those estimations to drive future pricing decisions. In general, one objective is to design a pricing algorithm that performs well in generating a small amount of worst-case regret.

Consequently, successful revenue management requires faithful product market value models and valid online statistical learning. Existing dynamic pricing studies focus on the faithfulness of adopted customer choice models (Myerson, 1981; Joskow and Wolfram, 2012; den Boer, 2015; Javanmard and Nazerzadeh, 2019; Mueller et al., 2019; Nambiar et al., 2019; Shah et al., 2019; Ban and Keskin, 2021; Javanmard et al., 2020), but, unfortunately, this is insufficient: certain iterates within their online optimization process may violate pre-specified optimization constraints (for example, sparsity constraint) and thus deny the validity of ultimate pricing decisions. Such lack of validity haunts practitioners' deployment of dynamic pricing systems and challenges scientists' craftsmanship: *how can one design an online regularization scheme to ensure the validity of online statistical learning uniformly among all decision points and secure low regret at the same time?* Specifically, we aim to deliver a regularization automation scheme based on learned-online market knowledge.

1.1 Our contributions

In this work, we make the decisive observation that the always-validity of pricing decisions builds on the *online regularization* scheme. This insight is drawn from an elegant interplay between sparse online statistical learning and non-asymptotic martingale concentration, which is desirable to establish the always-validity of pricing decisions. Such interplay leads us to propose a novel online regularization scheme: we identify uncertainties surrounding learned product demand parameters and regularize them to ensure the feasibility of iterating over all decision points within the pre-specified confidence budget. In such a sense, a successful always-valid high-dimensional dynamic pricing algorithm design will always return valid pricing decisions with high probability. Hence, we regularize sparse online statistical learning by quantifying and offsetting uncertainties evolving within the estimation process.

We call this principle technical tool *Optimistic Online LASSO* (`OOLASSO`): a novel *online regularization* scheme for online lasso. Based on it, we propose an optimistic online regularized

maximum likelihood pricing (OORMLP) algorithm. The OORMLP enjoys three major advantages: encode market noise knowledge into pricing process optimism; empower online statistical learning with always-validity overall decision points; envelop estimation error process with time-uniform non-asymptotic concentration bounds. These properties ensure the validity and robustness of our algorithm in practical dynamic pricing problems. In theory, we establish (OOLASSO) a non-asymptotic time-uniform oracle inequality of our estimator. Such inequality is possible by our novel use of non-asymptotic martingale concentration inequalities (Maillard, 2019; Howard et al., 2020) to ensure the always-validity warranty under a user-specified confidence budget. Built upon this time-uniform oracle inequality, we further show that our OORMLP algorithm achieves a logarithm regret bound, which meets the information-theoretical lower bound in the literature (Theorem 5.1, Javanmard and Nazerzadeh (2019)). In the experiment, we evaluate the performance of OORMLP in both synthetic and real data set. The results back up our theoretical superiority of OORMLP algorithm in its robustness perspective against different demand uncertainties. Besides, we demonstrate how OORMLP utilizes the user-specified confidence budget into an online regularization scheme to trade off price exploration and exploitation to achieve a substantial regret reduction in finite time performance compared to RMLP (Javanmard and Nazerzadeh (2019)).

In summary, our paper makes the following three major contributions.

1. Conceptually, we formulate the continuous monitoring of high-dimensional dynamic pricing problems. Our formulation bridges the high-dimension statistics literature in the Statistics community with continuous monitoring literature in the Operations Research community, opening a new venue for future studies on practical online statistical learning frameworks.
2. Methodologically, we propose the OORMLP algorithm for continuous monitor high-dimensional dynamic pricing to ensure the pricing strategy is valid at any time. To our knowledge, this is the first high-dimensional dynamic pricing algorithm with an always-valid guarantee.
3. Theoretically, we establish time-uniform Lasso oracle inequalities on the estimation error

process and further show a time-uniform logarithmic regret bound for our `OORMLP` algorithm. As a technical by-product, we develop `OOLASSO` to manage the optimism of online LASSO procedure via our novel use of non-asymptotic martingale concentration.

1.2 Related literature

Our work contributes to the learning-based dynamic pricing literature in problem formulation, to regularized online statistical learning in methodology, and to the growing literature of always valid online decision-making in theory.

Dynamic pricing with demand learning. Dynamic pricing with learning is a field of research that investigates pricing algorithms for situations when the demand function is unknown. Typically, the challenge is described as a form of the multiarmed bandit problem, with the arms being priced and the payoffs from the different arms being correlated, due to the measurements of demand assessed at different price points being correlated random variables. This includes parametric approaches (Broder and Rusmevichientong, 2012; Keskin and Zeevi, 2014; Broder and Rusmevichientong, 2012), semi-parametric ones (Shah et al., 2019) as well as nonparametric ones (Fan et al., 2021; Liu et al., 2022; Keskin and Zeevi, 2014). Beyond these studies, our work advances the problem formulation from finite to randomly stopped and possibly infinite horizon to meet the demand of continuous monitoring dynamic pricing in modern online service industrial practice.

A more related line of work is contextual dynamic pricing, which can be categorized into three groups, with different emphasis on how the context plays roles in the price and products market demand or value. The first group of references (Qiang and Bayati, 2016; Nambiar et al., 2019; Wang et al., 2021) uses context x as covariates of market demand. They assume the demand is observable and has a relationship with the offered price and the product context. In our work, we don't observe the demand but only the sale status of a product. The second group of references (Mao et al., 2018; Cohen et al., 2020) considers a noise-less

contextual dynamic pricing, which captures the relationship between value and product context in a deterministic way. The third group of references (Javanmard and Nazerzadeh, 2019; Luo et al., 2021; Fan et al., 2022) considers a noisy linear valuation model, which is also the model used in our paper.

Regularized online statistical learning. In the past decade, regularized offline statistical learning methodology, including Ridge regression (Hoerl and Kennard, 1970) and Lasso regression (Tibshirani, 1996) and related high-dimensional literature (Bühlmann and Van De Geer, 2011; Negahban et al., 2012; Wainwright, 2019), have found their applications integral to the solution for various online machine learning task. The applications span across several different tasks including bandit algorithms design (Wang et al., 2020a; Wu et al., 2022), online decision making (Bastani and Bayati, 2020; Wang and Cheng, 2020; Chen et al., 2021a,b; Wang and Li, 2022) and high-dimensional dynamic pricing (Javanmard and Nazerzadeh, 2019; Fan et al., 2021). Indeed, these efforts inspired people to several proof concepts and elegant statistical frameworks for online machine learning tasks. However, the associated calibration scheme for regularization level in these prior efforts is typically designed for offline uncertainty (where the dataset is assumed given) but not online uncertainty (where the dataset is not given), leading to concerns about the validity of online-learned models and the consequent inference result. Beyond these studies, our work advances the methodology of regularized statistical learning from a constant level regularization for offline uncertainty to a process level regularization for online uncertainty, which we term *online regularization*.

Such online regularization marks the key difference of our work compared to the `RMLP` in Javanmard and Nazerzadeh (2019), which also considered sparse learning in high-dimensional dynamic pricing. In practice, addressing the continuous monitoring high-dimensional dynamic pricing problems requires rethinking on the art of `RMLP` in the following three respects: (1) Rethink how to formulate the online uncertainty. In `RMLP`, the noise is assumed to be i.i.d, which does not capture the dependency nature between observations in the online setting.

In contrast, we consider a martingale difference noise distribution, which is more suitable to quantify online uncertainty. (2) Rethink the product feature sequence distribution. In `RMLP`, the product feature vectors are independently and identically sampled from a fixed distribution. In contrast, our framework allows a non-i.i.d. or general feature distribution. (3) Rethink the regularization level sequence. `RMLP` considers episode updates and requires resetting the algorithm. In contrast, our design of regularization does not need to reset the algorithm, hence is more sample efficient. Moreover, our regularization design mechanism also includes product context uncertainty and confidence budget to better balance the tradeoff between online uncertainty and online estimation error. See Remark 1 and Remark 5 for more detailed comparisons of these two methods.

Always-valid online decision making. Always-valid online decision making is an emerging field of studies in the last half decade (Johari et al., 2015; Zhao et al., 2016; Johari et al., 2021). Such emergence is a response of surging demand from modern online service industrial practice since the opportunity cost of lengthy online experiments is high and regrettable (Johari et al., 2021). Indeed, it is preferable to determine the real impact as fast as feasible or to terminate the ongoing experiment if the result looks unpromising, allowing the scientist to try other activities. Additionally, adjusting the runtime length in advance is unfeasible due to a lack of knowledge about the amount of the seeking impact and cost elasticity. Consequently, in modern online service practice, such dynamic trade-offs between greatest effect detection and shortest running time constrain the implementation of online statistical learning methodologies. Our work makes a first advance on the theory of always-valid online decision making into the high-dimensional dynamic pricing problems.

2 High-dimensional dynamic pricing problems

This section defines the high-dimensional dynamic pricing problems. Section 2.1 provides a five-step general design and essential elements of dynamic pricing algorithms. Section 2.2

provides our statistical framework for the market value of the product. Section 2.3 provides our presumption on the implemented pricing function.

2.1 A general design of dynamic pricing algorithms

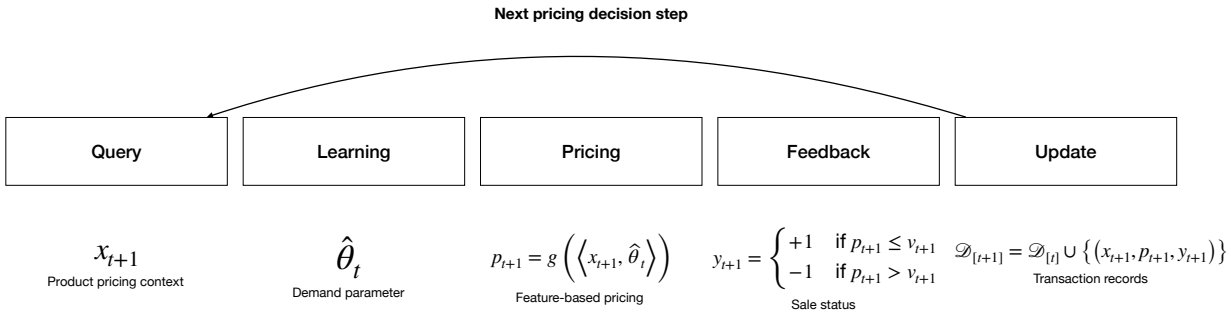

Figure 1: A general design of dynamic pricing algorithms

In a dynamic pricing problem with decision horizon T, the agent is required to determine total T prices at decision points $1, 2, \cdots, T$. Here T is an *unknown integer-valued random variable* and its realization is determined by an unknown terminating rule from a decision maker. At a decision point $t \in [T]$, a customer in the market selects a product with context x_t from a d-dimensional unit sphere $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_\infty \le 1\}$. The agent receives a pricing query for x_t, and her goal is to choose a posted price $p_t \in \mathbb{R}$ to maximize the revenue. The market value v_t of product x_t is unknown. After posting a price p_t, the customer decides whether to purchase the product based on market value v_t. The market value v_t is not observable to the agent, but only a binary-valued sale status variable $y_t \in \{-1, +1\}$. If $p_t \le v_t$, a sale occurs and the seller collects a revenue p_t and $y_t = +1$; otherwise, no sale occurs and no revenue is received and $y_t = -1$. Formally,

$$y_t = \begin{cases} +1 & \text{if } p_t \le v_t \\ -1 & \text{if } p_t > v_t \end{cases} \tag{1}$$

The seller's objective is to develop a pricing policy that maximizes revenue received.

Figure 1 briefly summarizes a general design of dynamic pricing algorithms for revenue

maximization via illustration of the five steps in a single decision step. In particular, at each decision point $t+1$, the agent

1. **Query**: The algorithm receives a query for pricing on the product with high-dimensional context vector $x_{t+1} \in \mathcal{X}$.
2. **Learning**: The algorithm learns a demand parameter estimate $\widehat{\theta}_t \in \Omega$ based on up-to-time t transaction records $\mathcal{D}_{[t]} = \{(x_s, p_s, y_s)\}_{s=1}^{t}$ to predict market value v_{t+1} of product x_{t+1}.
3. **Pricing**: The algorithm posts a revenue-maximizing price $p_{t+1} = g(\widehat{\theta}_t; x_{t+1})$ with a user-specified pricing function g.
4. **Feedback**: The algorithm receives a sale status y_{t+1}, based on the product's sale price p_{t+1}.
5. **Update**: The algorithm updates the transaction records $\mathcal{D}_{[t+1]} = \mathcal{D}_{[t]} \cup \{(x_{t+1}, p_{t+1}, y_{t+1})\}$.

Building upon the above general design of dynamic pricing algorithms, our goal is to provide an online statistical learning framework that fulfills three desiderata–sparse learning, always-validity, and revenue-maximization–that outlined in Section 3 to resolve high-dimensional dynamic pricing problems in continuous monitoring setting. The resulting dynamic pricing algorithms and the statistical learning framework are established in Section 4 and their formal fulfillment to the three desiderata are elaborated in Section 5.

2.2 Product market value model

Our statistical framework for market value v_t of product x_t consists of three parts: the market value model $v_t|x_t$, the target demand parameter θ_0 and the martingale difference noise process $\{\eta_t\}_{t=1}^{T}$. First, we model market value v_t of the product as a linear function of the observable product covariate x_t; formally

$$v_t = \langle \theta_0, x_t \rangle + \eta_t. \tag{2}$$

Second, the unknown parameter θ_0 is the target demand parameter that characterizes the demand profile of customers' behaviors. Parallel to high-dimensional dynamic pricing literature (Javanmard and Nazerzadeh, 2019), we consider a structured feasible parameters

set Ω in which θ_0 is high-dimensional and sparse; formally, for user-specified constants s_0 and W, the feasible parameters set Ω is defined as

$$\Omega = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \leq s_0, \|\theta\|_1 \leq W\}. \tag{3}$$

Third, the noise process $\{\eta_t\}_{t=1}^T$ in (2) accounts for unmeasured context and random noises. Notably, we consider a more general and practical dependent noise process drawn from a martingale difference sequence that is adapted to current transaction records. That is, with respect to the σ-field

$$\mathcal{H}_{t-1} = \sigma(x_1, p_1, y_1, \cdots, x_{t-1}, p_{t-1}, y_{t-1}, x_t, p_t) \tag{4}$$

generated by all transaction records before y_t is observed, the noise process η_t satisfies $\mathbb{E}[\eta_t|\mathcal{H}_{t-1}] = 0$ for all $t \in [T]$. Our dependent noise process relaxes the i.i.d. assumption considered in Javanmard and Nazerzadeh (2019). The conditional distribution of $\eta_t|\mathcal{H}_{t-1}$ is assumed to be log-concave in this paper. Many common probability distributions such as normal, logistic, uniform, exponential, Laplace, and bounded distributions are log-concave (Wellner, 2012). In particular, we define the 'steepness' of a function $F_{\eta_t|\mathcal{H}_{t-1}}(\cdot)$ as

$$u_{W,t} \equiv \sup_{|x|\leq 3W} \left\{\max\left\{\log' F_{\eta_t|\mathcal{H}_{t-1}}(x), -\log'\left(1 - F_{\eta_t|\mathcal{H}_{t-1}}(x)\right)\right\}\right\} \tag{5a}$$

and also define the 'flatness' of function $F_{\eta_t|\mathcal{H}_{t-1}}(\cdot)$ as

$$l_{W,t} \equiv \inf_{|x|\leq 3W} \left\{\min\left\{-\log'' F_{\eta_t|\mathcal{H}_{t-1}}(x), -\log''\left(1 - F_{\eta_t|\mathcal{H}_{t-1}}(x)\right)\right\}\right\}. \tag{5b}$$

In addition, we define the *maximal steepness* to be the constant $u_W = \max_{t\in[T]} u_{W,t}$ and the *minimal flatness* to be the constant $l_W = \min_{t\in[T]} l_{W,t}$.

The above statistical framework of product market value induces a probabilistic model for the sale status process $\{y_t\}_{t=1}^T$. The sale status process denotes a trajectory of customer transaction decisions with respect to the corresponding pricing sequence $\{p_t\}_{t=1}^T$ and product

sequence $\{x_t\}_{t=1}^T$. In particular, given the definition of sale status (1) and the market value model (2), the sale status process $\{y_t\}_{t=1}^T$ is generated from the following probabilistic model:

$$\mathbb{P}_{\theta_0}(y_t|\mathcal{H}_{t-1}) = \begin{cases} 1 - F_{\eta_t|\mathcal{H}_{t-1}}(p_t - \langle\theta_0, x_t\rangle) & \text{if } y_t = +1, \\ F_{\eta_t|\mathcal{H}_{t-1}}(p_t - \langle\theta_0, x_t\rangle) & \text{if } y_t = -1, \end{cases} \tag{6}$$

where $F_{\eta_t|\mathcal{H}_{t-1}}(\cdot)$ denotes the conditional distribution of noise η_t given $\mathcal{H}_{t-1}$.

2.3 Pricing function

Our framework allows a flexible pricing function g used at Step 3 of the pricing algorithm design (Figure 1). Such a feature is standard in industrial practice to provide flexible deployment of dynamic pricing algorithms (Johari et al., 2021). We assume the pricing function g is a L-Lipschitz continuous function for some Lipschitz constant $L \leq 1$, which is satisfied by the common pricing function choice in the literature, given in Example 2.1.

Example 2.1. *To maximize the expected revenue, it is shown in auction theory (Myerson, 1981; Javanmard and Nazerzadeh, 2019), the revenue-maximizing price* $p^*(x_t) = \arg\max_p\{p(1 - F_{\eta_t|\mathcal{H}_{t-1}}(p - \langle\theta_0, x_t\rangle))\}$. *The first order conditions says that the optimal posted price* $p_t^* = p^*(x_t)$ *satisfy*

$$p_t^* = \frac{1 - F_{\eta_t|\mathcal{H}_{t-1}}\left(p_t^* - \langle\theta_0, x_t\rangle\right)}{f_{\eta_t|\mathcal{H}_{t-1}}\left(p_t^* - \langle\theta_0, x_t\rangle\right)} = p_t^* - \langle\theta_0, x_t\rangle - \phi_t(p_t^* - \langle\theta_0, x_t\rangle)$$

by letting $\phi_t(v) \equiv v - \frac{1-F_{\eta_t|\mathcal{H}_{t-1}}(v)}{f_{\eta_t|\mathcal{H}_{t-1}}(v)}$. *That is,* $\langle\theta_0, x_t\rangle + \phi_t(p_t^* - \langle\theta_0, x_t\rangle) = 0$ *and hence* $p_t^* = \langle\theta_0, x_t\rangle + (\phi_t)^{-1}(-\langle\theta_0, x_t\rangle) = g_t(\langle\theta_0, x_t\rangle)$. *So the pricing function has the closed form*

$$g_t(v) \equiv v + (\phi_t)^{-1}(-v), \tag{7}$$

where $\phi_t(v) \equiv v - (1 - F_{\eta_t|\mathcal{H}_{t-1}}(v))/f_{\eta_t|\mathcal{H}_{t-1}}(v)$ *is known as a virtual valuation function. By Lemma S7.4, the pricing function* g_t *is 1-Lipschitz continuous.*

3 Evaluating dynamic pricing policy

In this section, we elaborate on what makes a good dynamic policy. Our goal is to design a pricing policy π that offers the price $p_t(\pi)$ for the product x_t in order to (i) **learn** the true

demand parameter θ_0 to inform seller about the underlying product market value model (2), (ii) **continuously monitor** the estimation error of the estimated demand parameter, and (iii) **optimize** the posted price to maximize the expected revenue. In order for the policy π to fulfill the learning and optimizing tasks, it must satisfy the following desiderata: (A) it should return a sparse demand parameter estimate to enhance the explainability of the pricing mechanism and product market value, (B) it should be able to *adapt the online uncertainty* of product market value model (2) to obtain *always-valid* statistical error bounds, and (C) it should be *revenue-maximized*, i.e., the difference between posted price $p_t(\pi)$ and the oracle price π_t^* should be small. Consequently, it's critical to establish an effective strategy that strikes a balance between exploration (gathering data for learning parameters) and exploitation (offering optimal pricing based on learned parameters).

Having outlined the desiderata for our sought-after pricing policy, we now propose three properties of the online statistical learning framework that should be encoded in the adopted pricing policy. These properties are:

(A) Sparse Learning: the learned demand parameter identifies the subset of decisive pricing features to enhance the explainability of the learned market value model. (Section 3.1)

(B) Always-Validity: the estimation error of the online learned market value model remains statistical validity even when the pricing algorithm is terminated randomly. (Section 3.2)

(C) Revenue-Maximization: the collective revenue is comparable to the revenue of the oracle pricing policy which knows the true demand parameter. (Section 3.3)

3.1 Online Lasso procedure towards sparse learning

To achieve the first desiderata on learning sparse demand parameter estimate, we adopt the online Lasso procedure, defined as follows.

Definition 1. *We define the* ***online Lasso procedure*** *as follows:*

1. *At a decision point t, the agent calculates the negative log-likelihood function $\mathcal{L}(\theta; \mathcal{D}_{[t]})$ of a model parameter θ and up-to-time t transaction records $\mathcal{D}_{[t]}$ as*

$$\mathcal{L}_t(\theta) \equiv \mathcal{L}(\theta; \mathcal{D}_{[t]}) = t^{-1} \sum_{s=1}^{t} \log(1/\mathbb{P}_\theta(y_s|\mathcal{H}_{s-1})). \tag{8a}$$

The probability $\mathbb{P}_\theta(y_s|\mathcal{H}_{s-1})$ is from the Bernoulli model (6) of the sale status process $\{y_t\}_{t=1}^T$; that is, with $u_t(\theta) \equiv p_t - \langle \theta, x_t \rangle$, $\log(1/\mathbb{P}_\theta(y_s|\mathcal{H}_{s-1})) = \mathbb{I}(y_t = 1) \log\left(1/(1 - F_{\eta_t|\mathcal{H}_{t-1}}(u_t(\theta)))\right) + \mathbb{I}(y_t = -1) \log\left(1/F_{\eta_t|\mathcal{H}_{t-1}}(u_t(\theta))\right)$.

2. *The algorithm penalizes the loss $\mathcal{L}_t(\theta)$ by the l_1-norm penalty at regularization level $\lambda_t > 0$. In particular, at decision point t, the algorithm learns an estimator $\widehat{\theta}_t$ by solving the ℓ_1-regularized quadratic program*

$$\widehat{\theta}_t \equiv \arg \min_{\|\theta\|_1 \leq W} \left\{ \mathcal{L}_t(\theta) + \lambda_t \|\theta\|_1 \right\}. \tag{8b}$$

3. *Repeating the above Lasso procedure at each decision point $t = 1, 2, \cdots, T$, with a regularization level sequence $\{\lambda_t\}_{t=1}^T$, the agent thus learns at the decision horizon T an estimation sequence: $\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_T$.*

The online Lasso procedure (Definition 1) delivers a statistical learning framework for online sparse learning. In practice, given a regularization level sequence $\{\lambda_t\}_{t=1}^T$, the online Lasso procedure returns a sequence of constrained estimators $\{\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_T\}$ towards learning a sparse demand parameter estimate in the product market value model (2). Indeed, such online Lasso procedures benefit the interpretability of the resulting product market value model and the explainability of the pricing mechanism.

However, the benefit of the online Lasso procedure may be blocked by an improper choice of regularization level sequences $\{\lambda_t\}_{t=1}^T$. As well-recognized in the high-dimensional statistics literature, different regularization level sequences $\{\lambda_t\}_{t=1}^T$ lead to different properties of resulting constrained estimators sequence $\{\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_T\}$. As far as the continuous monitoring dynamic pricing concerns, the fundamental challenge is how to choose the regularization level λ_t in Lasso program (8b) at a process level, i.e. for every decision step t

from 1 to the random decision horizon T. Section 5.1 contributes the key observation that the online Lasso procedure builds and thrives on online regularization scheme design to calibrate online uncertainty during the pricing process.

3.2 Always valid estimation error bound process

To achieve the second desiderata on always-valid online statistical learning, we introduce the concept of always-valid estimation error bound process, defined as follows :

Definition 2. *Given any (possible unbounded) stopping time T with respect to historical filtration $\{\mathcal{H}_t\}_{t=0}^T$ (defined at (4)). A sequence of constant real number $\{r_t\}_{t=1}^T$ is an* ***always valid estimation error bound process*** *of the estimator sequence $\{\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_T\}$ with confidence budget α if it holds that*

$$\mathbb{P}_{\theta_0}\left(\exists t \in [T] : \|\widehat{\theta}_t - \theta_0\|_2 > r_t\right) \leq \alpha. \tag{9}$$

The always valid estimation error bound process (Definition 2) serves as a principal theoretical tool for online service industrial practice in the continuously monitoring risk control of adopted online statistical learning procedures. For an online learned estimator sequence $\{\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_T\}$, the corresponding error bound process $\{r_1, r_2, \cdots, r_T\}$ collectively gives a time-uniform control on the estimation error sequence $\{\|\widehat{\theta}_1 - \theta_0\|_2, \|\widehat{\theta}_2 - \theta_0\|_2, \cdots, \|\widehat{\theta}_T - \theta_0\|_2\}$ such that the probability of out-of-control is at most at the level of user pre-specified confidence budget α. Such time-uniform risk control allows users to terminate the dynamic pricing algorithm whenever they wish, *and* the result still maintains statistical validity.

Establishing such an always valid error-bound process, however, is technically challenging and far from understood in the literature. The reason is that, while estimation error bound result for fixed sample size Lasso regression had been systematically studied in the literature and inspired people for an elegant theoretical framework, they focused on offline uncertainty (the whole dataset is given) instead of online uncertainty (the dataset is not given and

is observed on the fly). Consequently, the classical method of high-dimensional statistics literature fails to meet the challenge of online statistical learning with continuous monitoring demanded in the modern online service industry. Section 5.2 contributes a key theoretical result on the always validity of online Lasso procedure (Definition 1).

3.3 Regret of a dynamic pricing policy

To achieve the third desiderata of revenue maximization, we define the notion of regret.

Definition 3. *The regret of a dynamic pricing policy π up to decision T is defined as*

$$\textbf{\textit{Regret}}_{\pi}(T) \equiv \max_{\theta_0 \in \Omega} \mathbb{E}\Big[\sum_{t=1}^{T} \big(r_t(p_t^*) - r_t(p_t(\pi))\big)\Big], \tag{10}$$

where $r_t(p) \equiv pI(v_t \geq p)$ is the expected revenue of the product x_t with the posted price p. The expectation is taken with respect to the noise η_t and product context x_t, and $p_t(\pi)$ denotes the price offered at decision step t by following policy π.

Definition 3 benchmarks the performance of a dynamic pricing policy π that determines posted prices $\{p_t\}_{t=1}^T$ to the corresponding 'oracle pricing policy', which exploits knowledge of the true demand parameter θ_0 and proposes the price $p_t^* = g(\langle \theta_0, x_t \rangle)$ for the product of context x_t, where $g(\cdot)$ is a user-specified pricing function. In Example 2.1, the optimal price p_t^* is the price that maximizes the expected revenue. Formally, we consider the goal of maximizing revenue as minimizing the maximum regret at Definition 3. As pursued as the third desiderata of pricing policy, the goal is to design an online statistical learning procedure such that the regret (10) is small.

4 The `OORMLP` algorithm and `OOLASSO` procedure

This section establishes our pricing policy design that achieves the three desiderata discussed in Section 3. We first propose the Optimistic Online Regularized Maximum Likelihood Pricing (`OORMLP`) algorithm (Algorithm 1) as the desirable dynamic pricing policy at Section

4.1. Then we elaborate our novel Optimistic Online Lasso procedure (OOLASSO) towards always valid online statistical learning at Section 4.2.

4.1 OORMLP algorithm

In this section, we present the proposed dynamic pricing policy at Algorithm 1. The presentation follows the general design of dynamic pricing algorithms in Figure 1. In particular, at decision point $t+1$, the agent learns the demand parameter estimator $\widehat{\theta}_t$ based on the current transaction records $\mathcal{D}_{[t]}$ via Lasso regression in (8b) at regularization level λ_t specified in the optimistic online regularization scheme (13). In addition, both the sample covariance matrix $\widehat{\Sigma}_{[t]}$ and the online regularization sequence $\{\lambda_t\}_{t=1}^T$ in (13) can be incrementally updated: at each decision point t,

$$\widehat{\Sigma}_{[t]} \leftarrow t^{-1}\left[(t-1)\widehat{\Sigma}_{[t-1]} + x_t x_t^\top\right]; \quad \lambda_t \leftarrow \lambda_{t-1}\sqrt{(1-t^{-1})\|\widehat{\Sigma}_{[t]}\|_\infty / \|\widehat{\Sigma}_{[t-1]}\|_\infty}.$$

Such property allows an efficient online implementation in the experiments.

4.2 Optimistic online lasso procedure

Here, we elaborate our novel approach to construct a learning process $\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_T$ for the target demand parameter θ_0 based on transaction records $\mathcal{D}_{[t]} = \{(x_s, p_s, y_s)\}_{s=1}^t$ with optimism in the face of online uncertainty during the pricing process.

Definition 4. *An online Lasso procedure (Definition 1) is* ***optimistic*** *if the regularization sequence $\{\lambda_t\}_{t=1}^T$ is specified by the following* ***optimistic online regularization scheme****:*

$$\lambda_t(\alpha) \equiv 4u_W\sqrt{2\cdot t^{-1}\|\mathit{diag}(\widehat{\Sigma}_{[t]})\|_\infty \ln(2d/\alpha)}. \tag{13}$$

Definition 4 presents our novel regularization scheme for regulating online uncertainty during the dynamic pricing process. The reason we call (13) optimistic is that it regularizes the online LASSO procedure with optimism in the face of both demand uncertainty and product feature uncertainty during the dynamic pricing process, given a specified confidence

Algorithm 1 Optimistic Online Regularized Maximum Likelihood Pricing (`OORMLP`)

Require: Steepness of market noise u_W, pricing function $g(\cdot)$ and confidence budget α.

Initialization: Receive product context x_1. Post price p_1. Receive sale status y_1.

$\mathcal{D}_{[1]} \leftarrow \{(x_1, p_1, y_1)\}$; $\widehat{\Sigma}_{[1]} \leftarrow x_1 x_1^\top$; $\lambda_1 \leftarrow 4u_W\sqrt{2\|\mathrm{diag}(\widehat{\Sigma}_{[1]})\|_\infty \ln(2d/\alpha)}$.

for $t = 2, \ldots, [T]$ **do**

1.Query: Receive product context x_t.

2.Learning: Update the sample covariance matrix and regularization level:

$$\widehat{\Sigma}_{[t]} \leftarrow t^{-1}\left[(t-1)\widehat{\Sigma}_{[t-1]} + x_t x_t^\top\right], \tag{11a}$$

$$\lambda_t \leftarrow \lambda_{t-1}\sqrt{(1-t^{-1})\|\widehat{\Sigma}_{[t]}\|_\infty / \|\widehat{\Sigma}_{[t-1]}\|_\infty}; \tag{11b}$$

Update the estimate

$$\widehat{\theta}_{t-1} \leftarrow \arg\min_{\|\theta\|_1 \le W} \{\mathcal{L}_{t-1}(\theta) + \lambda_{t-1}\|\theta\|_1\}. \tag{12}$$

3.Pricing: Post price $p_t \leftarrow g\left(\left\langle \widehat{\theta}_{t-1}, x_t \right\rangle\right)$.

4.Feedback: Receive sale status y_t.

5.Update: $\mathcal{D}_{[t]} \leftarrow \mathcal{D}_{[t-1]} \cup \{(x_t, p_t, y_t)\}$.

end for

budget α. Three factors contribute to the regularization level $\lambda_t(\alpha)$. First, the constant u_W is the *maximal steepness* of noise process (5a) and represents our prior knowledge of demand uncertainty. Second, the empirical covariance matrix $\widehat{\Sigma}_{[t]} = t^{-1}\sum_{s=1}^{t} x_s x_s^\top$ characterizes the uncertainty of up-to-now product context sequence. Third, the constant α stands for the user-pre-specified confidence budget for the always-validity of implemented online LASSO procedure. These factors collectively express optimism in the face of online uncertainty during the dynamic pricing process and are the foundation to fulfill the three desiderata we pursued in Section 3. Consequently, we adopt the optimistic online regularization scheme (13) to design `OORMLP` algorithm (Algorithm 1) to enjoy three desiderata-sparse learning, always-validity and revenue-maximization-on resulting dynamic pricing policy.

Remark 1. *(Regularization comparison to `RMLP` in Javanmard and Nazerzadeh (2019)) The*

relation between our regularization scheme and the one in ***RMLP*** *is*

$$\lambda_{t,\texttt{OORMLP}}(\alpha) = \lambda_{t,\texttt{RMLP}}\sqrt{2\frac{\log_2(t)}{t}}\sqrt{\frac{\log(2d/\alpha)}{\log(d)}}\sqrt{\|diag(\widehat{\Sigma}_{[t]})\|_\infty}.$$

The relation above indicates that, while ***RMLP*** *do not,* ***OORMLP*** *includes in the regularization level the uncertainty arising from context sequence ($\|diag(\widehat{\Sigma}_{[t]})\|_\infty$). The root reason why* ***RMLP*** *does not take $\|diag(\widehat{\Sigma}_{[t]})\|_\infty$ into their regularization design is due to their assumption on the independent identical distributed property on the product sequence. When the distribution of product sequence deviates from such i.i.d. assumption, the regularization level in* ***RMLP*** *is improper to account for the effective noise process in the LASSO procedure. Our regularization design takes the context sequence uncertainty into account, which leads to the robustness of* ***OORMLP*** *in the experiments.*

5 Always-validity and regret analysis

This section elaborates on formal guarantees of three qualities of our `OORMLP` algorithm and `OOLASSO` procedure. Section 5.1 demystifies the design principle behind our optimistic online regularization scheme, formally achieving the first desiderata: sparse learning. Section 5.2 establishes the time-uniform Lasso oracle inequality (Theorem 1), formally achieving the second desiderata: always-validity. Section 5.3 present regret analysis (Theorem 2) of our `OORMLP` pricing policy, formally achieving the third desiderata: revenue-maximizing.

5.1 Optimistic online regularization scheme

This section demystifies the optimistic online regularization scheme (13) as a formal guarantee of the sparse learning of our online statistical learning framework.

5.1.1 Basic design principle

We now explain the design principle of the regularization sequence $\{\lambda_t\}_{t=1}^T$ for the optimistic online regularization scheme at (13). In principle, our goal is to design a regularization

sequence $\{\lambda_t\}_{t=1}^T$ that warrants the online LASSO procedure (Definition 1) with always-validity by constructing an always valid estimation error bound process (Definition 2). Intuitively, the optimal choice of the sequence is an outcome of the bias-and-variance trade-off. Bias arises as a shrinkage effect from l_1-regularizer and grows as λ_t increases. Besides, l_1-regularizer offsets fluctuations in the score function process $\{\nabla \mathcal{L}_t(\theta)\}_{t=1}^T$. Hence, an optimal choice of $\{\lambda_t\}_{t=1}^T$ is the smallest *envelop* that is large enough and *always* controls score fluctuations during the whole pricing process.

To obtain an always valid estimation error bound process of the online LASSO procedure (8b), we generalize standard guidance from high-dimensional statistics literature to the *process* level by considering the event

$$\mathfrak{G}(\{\lambda_t\}_{t=1}^T) = \left\{\forall t \in [T] : 4t^{-1}\|\nabla \mathcal{L}_t(\theta_0)\|_\infty \leq \lambda_t\right\}. \tag{14}$$

Given the above event, Theorem 1 in Section 5.2 shows that it is possible to build an *always valid* estimation error bound on the proposed online LASSO procedure. Therefore, an optimal design of $\{\lambda_t\}_{t=1}^T$ should be the one to ensure that $\mathfrak{G}(\{\lambda_t\}_{t=1}^T)$ holds with high probability.

Toward finding such an optimal selection, for a given confidence budget $\alpha \in (0, 1)$, our goal is to find a regularization level sequence $\{\lambda_t(\alpha)\}_{t=1}^T$ that satisfies

$$\mathbb{P}_{\theta_0}\left(\mathfrak{G}(\{\lambda_t(\alpha)\}_{t=1}^T)\right) \geq 1 - \alpha. \tag{15}$$

As supported by Lemma 1 in Section 5.1.2, the proposed optimistic online regularization scheme (Definition 4) satisfies the property (15). Therefore, when the agent learns the target demand parameter θ_0 by solving the LASSO problem in (8b) with the specified optimistic online regularization scheme in (13), the resulting estimator process $\{\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_T\}$ enjoys an *always-validity*, i.e., the implemented online statistical learning procedure is theoretically valid at each decision point with a time-uniform estimation error bound (Theorem 1). Such always-validity serves as a warranty on the robustness and safety of dynamic pricing algorithm design and fulfills the second desiderata pursued in Section 3.

5.1.2 Formal design

Here, we give a formal derivation of the online regularization scheme design to implement the principle outlined in Section 5.1.1. To analyze the event of valid Lasso procedure (14), we first show a consequence of optimistic online regularization scheme (13) on the infinity norm of score function process:

Lemma 1. *(Always Valid Score Function Process Bound) Under the optimistic online regularization scheme (13), it holds with probability at least* $1-\alpha$ *that*

$$\forall t \in [T] : \|\nabla \mathcal{L}_t(\theta_0)\|_\infty \leq u_W \sqrt{2t^{-1}\|diag(\widehat{\Sigma}_{[t]})\|_\infty \ln(2d/\alpha)}. \tag{16}$$

Lemma 1 provides a time-uniform control on the score function process $\{\nabla \mathcal{L}_t(\theta_0)\}_{t=1}^T$ of their infinity norm process. Concretely, the result bounds the fluctuation of score function process $\{\|\nabla \mathcal{L}_t(\theta_0)\|_\infty\}_{t=1}^T$ at the true demand parameter θ_0 by carefully designing the online regularization sequence $\{\lambda_t\}_{t=1}^T$ to adaptive realized online uncertainty at each decision point. As remarked in Section 5.1, the online regularization scheme (13) warrants always-validity of the `OOLASSO` procedure. Consequently, the design of optimistic online regularization scheme (13) follows from Lemma 1 and the event of valid Lasso procedure (14).

Remark 2. *An advantage of the always-valid type result in Lemma 1 is that it holds for not only a constant decision horizon* T *(independent from the pricing process) but also a random decision horizon* $T(w)$ *(dependent on the pricing process). This property enables us to do valid inferences at randomly stopped times.*

5.1.3 Exploration-exploitation trade-off

We briefly discuss how the proposed optimistic online regularization scheme (13) balances the explore-exploit trade-off during the pricing process. As we will show in Theorems 1 and 2, the revenue loss of the `OORMLP` in each decision point t is of the same order as the squared estimation error bound $\|\widehat{\theta}_t - \theta_0\|_2^2$ which is bounded by λ_t^2. Thus, the regularization level λ_t

determines the pricing optimism of `OORMLP`. Price with larger revenue loss can be viewed as "price exploration" since larger price uncertainty helps the learning of θ_0. On the other hand, a price with a smaller revenue loss can be viewed as "price exploitation", indicating that the agent exploits the learned demand parameter to maximize the collected revenue.

In general, the proposed optimistic online regularization scheme (13) delivers a pricing policy that gradually shifts from price exploration to price exploitation. There are three main factors that contributed to pricing optimism: market noise knowledge u_W, product context process $\widehat{\Sigma}_{[t]}$, and confidence budge α. Each of them captures different uncertainties happening in dynamic pricing, where u_W measures demand uncertainty, $\widehat{\Sigma}_{[t]}$ measures product feature uncertainty and α measures online procedure uncertainty. Section 5.1 explains how these factors contribute to the regularization level in the face of online uncertainty. Section 6 investigates how these factors contribute to pricing optimism in the numerical experiments.

5.2 Time-uniform lasso oracle inequality

This section establishes the time-uniform lasso oracle inequality (Theorem 1) as a formal guarantee of the always-validity (the second desiderata; Section 3.2) of our framework.

To derive an error envelop for the estimates $\{\widehat{\theta}_t\}_{t=1}^T$ produced from `OOLASSO`, we first define a restricted eigenvalue process condition as a process analogue of a standard requirement in high-dimensional statistical estimation (Wainwright, 2019).

Definition 5. *For a product context process $\{x_t\}_{t=1}^T$, we say it satisfies a **restricted eigenvalue process** condition if there exists a sequence of positive number $\{\phi_t^2\}_{t=1}^T$ such that*

$$\forall t \in [T] : \min_{J\subseteq[d];|J|\leq s_0} \min_{v\neq 0;\|v_{J^c}\|_1\leq 3\|v_J\|_1} \left(v^\top \widehat{\Sigma}_{[t]} v\right) / \|v\|_2^2 \geq \phi_t^2, \tag{17}$$

where v_J is the vector obtained by setting the elements of v that are not in J to zero.

Remark 3. *(On the requirement of product context sequence $\{x_t\}_{t=1}^T$) Here, we only present the widely adopted restricted eigenvalue condition on the product context sequence $\{x_t\}_{t=1}^T$ to*

prove the time-uniform oracle inequality. Such conditions on the product context sequence can be relaxed by adapting arguments in high-dimensional inference literature (See, for example, Chichignoud et al. (2016)).

Remark 4. *(On the lower bound sequence $\{\phi_t^2\}_{t=1}^T$) Let Σ_0 be the population covariance matrix of product context x_t and denote its restricted eigenvalue as $\phi^2(\Sigma_0, s_0)$. Based on matrix martingale concentration arguments, it can be shown that a choice of the lower bound sequence $\{\phi_t^2\}_{t=1}^T$ under confidence budget α is*

$$\phi_t^2 = \phi^2(\Sigma_0, s_0) - 32 s_0 \left[\sqrt{2t^{-1} \ln(d(d+1)/2\alpha)} + t^{-1} \ln(d(d+1)/2\alpha) \right].$$

Theorem 1. *(Always valid estimation error bound process) Suppose the product contexts process $\{x_t\}_{t=1}^T$ satisfies the restricted eigenvalue condition (17) with a non-random sequence $\{\phi_t^2\}_{t=1}^T$. Then, under the online regularization scheme (13), it holds that:*

$$\mathbb{P}_{\theta_0} \left(\exists t \in [T] : \left\| \widehat{\theta}_t - \theta_0 \right\|_2^2 \geq \frac{16 s_0 \lambda_t^2(\alpha)}{l_W^2 \phi_t^2} \right) \leq \alpha. \tag{18}$$

Theorem 1 provides a formal guarantee of the always-validity of our online statistical learning framework. With a such guarantee, the user is allowed to terminate the dynamic pricing algorithm whenever they wish, and the result of estimation error bound maintains statistical validity. In particular, Theorem 1 indicates that the convergence rate of learning demand parameter θ_0 is determined by three primary factors: (1) Non-smoothness of martingale difference noise conditional distribution function $F_{\eta_t|\mathcal{H}_{t-1}}$. This is captured by the minimal flatness defined by (5b). It controls the amount of information about the mean market value $\langle x_t, \theta_0 \rangle$ of product x_t at each time step t. (2) The rate at which the product context x_t explores the parameter space. This is governed by the restricted eigenvalue process condition (Definition 5). If the lower bound sequence $\{\phi_t^2\}_{t=1}^T$ is small, the product context is relatively aligned and one requires a larger sample size to estimate the demand parameter within a specified accuracy. (3) The complexity of demand parameter θ_0. This is captured through the sparsity measure s_0 in the feasible parameter space (3).

5.3 Regret analysis of the OORMLP algorithm

This section establishes the regret analysis (Theorem 2) of the proposed OORMLP dynamic pricing algorithm (Algorithm 1) as a formal guarantee of the revenue-maximization quality (the third desiderata; Section 3.3) of our online statistical learning framework. The following theorem bounds the regret of the proposed OORMLP dynamic pricing algorithm.

Theorem 2. *(Regret guarantee for OORMLP algorithm) Suppose the product context sequence* $\{x_t\}_{t=1}^T$ *satisfies the restricted eigenvalue condition* (17) *with a non-random sequence* $\{\phi_t^2\}_{t=1}^T$. *Then, under the online regularization scheme* (13), *with probability at least* $1-\alpha$,

$$\textbf{\textit{Regret}}_{\texttt{OORMLP}}(T) \leq \frac{256 C s_0 u_W^2}{l_W^2 \min_{t\in[T]} \phi_t^2} \ln(\frac{2d}{\alpha}) \log T. \tag{19}$$

To read the regret bound (19), we break it into three elements of dynamic pricing problems. First, the regret bound depends on the product market value model (Sec. 2.2) in terms of s_0, the sparsity level of demand coefficient, and d, the dimension of product context, at the rate $s_0 \log d$. Second, the regret bound depends on the martingale difference noise process $\{\eta_t\}_{t=1}^T$ at (2) in terms of u_W, the maximal steepness and l_W, the minimal flatness, at the rate $(u_W/l_W)^2$. Third, the regret bound depends on the product context sequence $\{x_t\}_{t=1}^T$ via the restricted eigenvalue sequence $\{\phi_t^2\}_{t=1}^T$ (Definition 17) at the rate $1/\min_{t\in[T]} \phi_t^2$. Under additional assumptions on the boundedness of these parameters, we achieve an $O(\log T)$ regret bound of OORMLP, which meets the information-theoretical lower bound shown in Javanmard and Nazerzadeh (2019).

Remark 5. *(Comparison to RMLP algorithm proposed in Javanmard and Nazerzadeh (2019))* *We emphasize that our regret bound (Theorem 2) is always valid in the sense that the result holds for a random decision horizon* T. *In contrast, the regret bound of RMLP only holds for a fixed constant decision horizon* T. *This is because the RMLP algorithm used the doubling trick to apply batch-type concentration result based on i.i.d. noise assumption in dynamic pricing algorithm design, while our result is based on martingale concentration. First, RMLP is not as*

sample efficient as ***OORMLP****. This is because* ***RMLP*** *needs to reset the algorithm several times during the pricing process to achieve logarithm regret. On the other hand, our* ***OORMLP*** *uses a novel non-asymptotic martingale concentration to avoid resetting the algorithm during the whole pricing process and still achieves logarithm regret. Second,* ***RMLP*** *relies on an i.i.d. noise assumption, while* ***OORMLP*** *allows for a more flexible martingale difference noise. As shown in experiments in Section 6,* ***OORMLP*** *is more sample efficient and robust to noise assumptions.*

6 Experiments

We evaluate the performance of the proposed OORMLP algorithm on both synthetic and real-world data. Additional simulations with dependent context sequence, sensitivity tests are provided in the supplement.

6.1 Simulations with independent context sequence

We compare OORMLP with RMLP under four representative demand uncertainty settings: (i) Gaussian ($\eta_t \sim N(0,1)$) (ii) Laplace ($\eta_t \sim \text{Laplace}(0,1)$) (iii) Periodic ($\eta_t = \sin(0.01t)$) and (iv) Cauchy ($\eta_t \sim \text{Cauchy}(0,1)$). Settings (i) and (ii) stand for instances of log-concave distributions, where (ii) has a heavier tail than (i). Setting (iii) stands for an instance of time-series noise, where the noises between two adjacent time points are strongly dependent. Setting (iv) stands for distribution beyond the log-concave distribution assumed in our theoretical analysis. This setting investigates our algorithm under model misspecification. We set $\theta_0 = (1,1,1,0,0,0,0,0,0,0)$ with $d = 10$. Each entry in the product context vector $x_t \in \mathbb{R}^{10}$ is generated from $N(0,1)$ and truncated to $[-1,1]$ (Similar synthetic data generation procedure is implemented by Bastani and Bayati (2020)). Therefore, $\|x_t\|_\infty \leq 1$.

We implement our OORMLP algorithm at two confidence budgets ($\alpha = 0.05$ and 0.1) which refer to different levels of pricing optimism, and compare our results with RMLP in Javanmard and Nazerzadeh (2019). In real scenarios, we do not know the exact distribution of demand

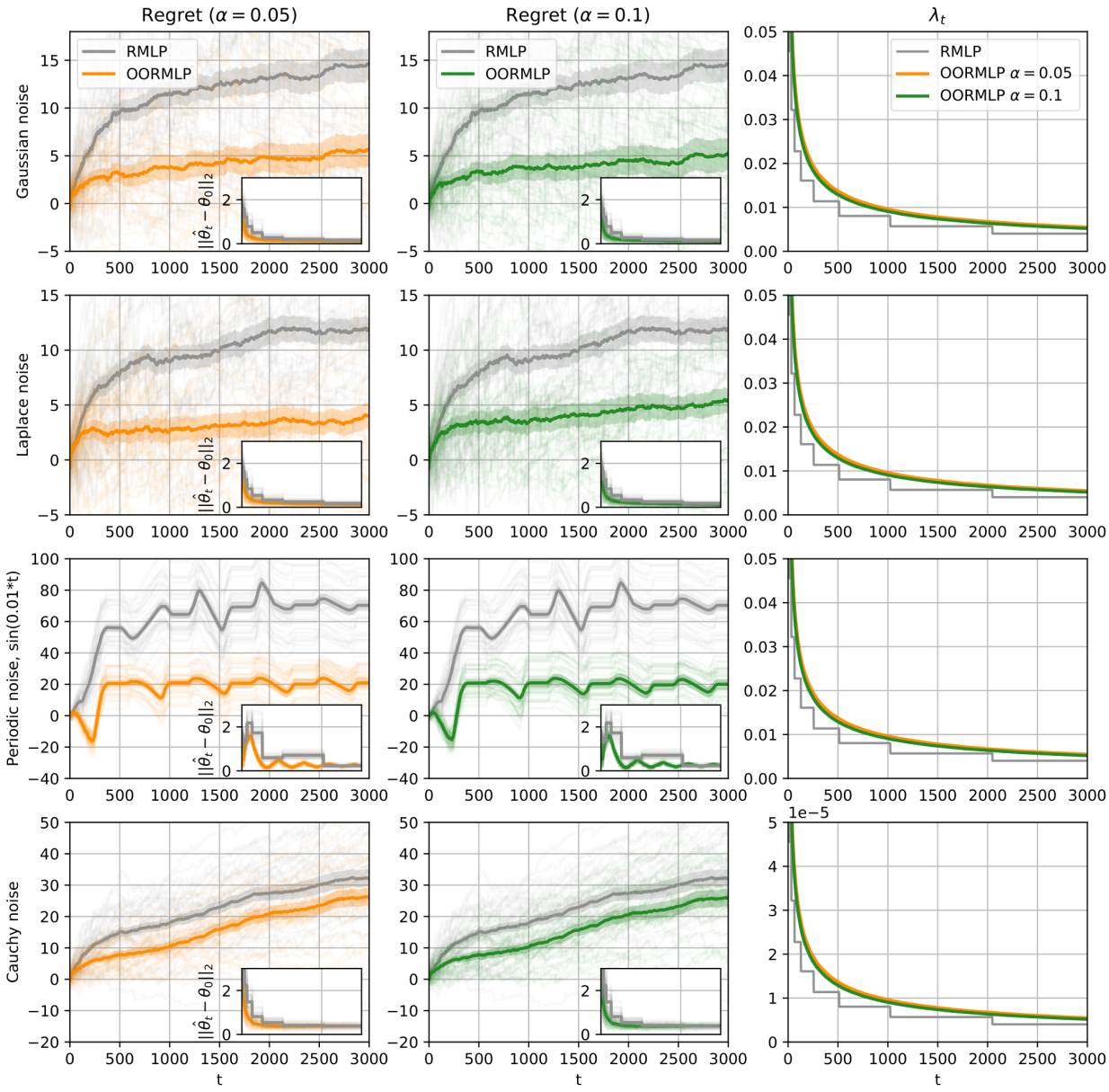

Figure 2: Comparison between `RMLP` and `OORMLP` when $d = 10$. **First row:** $\eta_t \sim N(0, 1)$. **Second row:** $\eta_t \sim \text{Laplace}(0, 1)$. **Third row:** $\eta_t = \sin(0.01t)$. **Fourth row:** $\eta_t \sim \text{Cauchy}(0, 1)$. **Two columns on the left:** different choices of confidence budget α. **Rightmost column:** λ_t for the experiments. **Small figures in each subfigure:** Estimation error $\|\widehat{\theta}_t - \theta_0\|_2$. Each transparent line represents one experiment. The solid lines and error bars represent the sample mean and its standard deviation.

uncertainty in advance, and hence we design the pricing function $g(\cdot)$ by assuming the uncertainty is standard normal ($\eta_t \sim N(0, 1)$). Such consideration tests the robustness of our algorithm when the demand uncertainty is unknown. Since $\|\theta_0\|_1 = 3$, we set $W = 10$ for

both OORMLP and RMLP. In practice, the theoretical online regularization choice in (13) might be conservative. To compare the finite-time performance of OORMLP and RMLP, we scale the regularization sequence $\{\lambda_t\}_{t=1}^T$ of both methods by the same scaling parameter $c_\lambda = 0.001$ (except for the Cauchy noise setting where we use $c_\lambda = 10^{-6}$ for both methods). We compute the mean and confidence interval of regrets over 32 replications. Figure 2 reports the results for the regret, the estimation error, and the regularization sequence used, which show the superiority and robustness of our algorithm (See Remark 1).

Below we give general remarks and rationales of our OORMLP from the perspectives of variance control, sample efficiency, and regret reduction.

1. **Sample efficiency on estimation error process.** Small figures in each subfigure at Figure 2 visualize the estimator error process of RMLP and OORMLP. In the first three uncertainty settings, OORMLP achieves smaller estimation errors than RMLP. This aligns with Remark 5 that OORMLP is more sample efficient than RMLP since it avoids resetting the algorithm. Remarkably, the estimator accuracy of RMLP is especially fragile in the setting (iii) of periodic noise. This is because RMLP uses samples only from previous episodes and updates geometrically, and its estimation accuracy and pricing performance are impeded in a scenario where noises between two adjacent time points are strongly dependent. In contrast to RMLP, our OORMLP enjoys a superior design in terms of sample efficiency and robustness in such periodic noise settings. Finally, in setting (iv) of Cauchy noise which violates our log-concave noise assumption, OORMLP performs similarly to RMLP.

2. **Confidence budget and regret reduction.** Similar to the performance in the estimation error process, OORMLP achieves much smaller regrets than RMLP in the first three uncertainty settings. The first two columns in the first row of Figure 2 show an interesting phenomenon that a larger confidence budget α leads to a more substantial regret reduction of our OORMLP, while the performance of RMLP is not adaptive to α. This aligns with our discussion in Section 5.1 on how OORMLP balances the explore-exploit trade-off during the pricing process.

3. **Shape of online regularization scheme.** The rightmost column of Figure 2 visualizes how non-asymptotic martingale concentration arguments authorize a process-level online regularization scheme. Compared to `RMLP` which resets itself geometrically (when $t = 2^k, k \in \mathbb{N}$) without considering product feature uncertainty, our `OORMLP` delivers a smooth regularization process against both product context uncertainty and demand uncertainty.

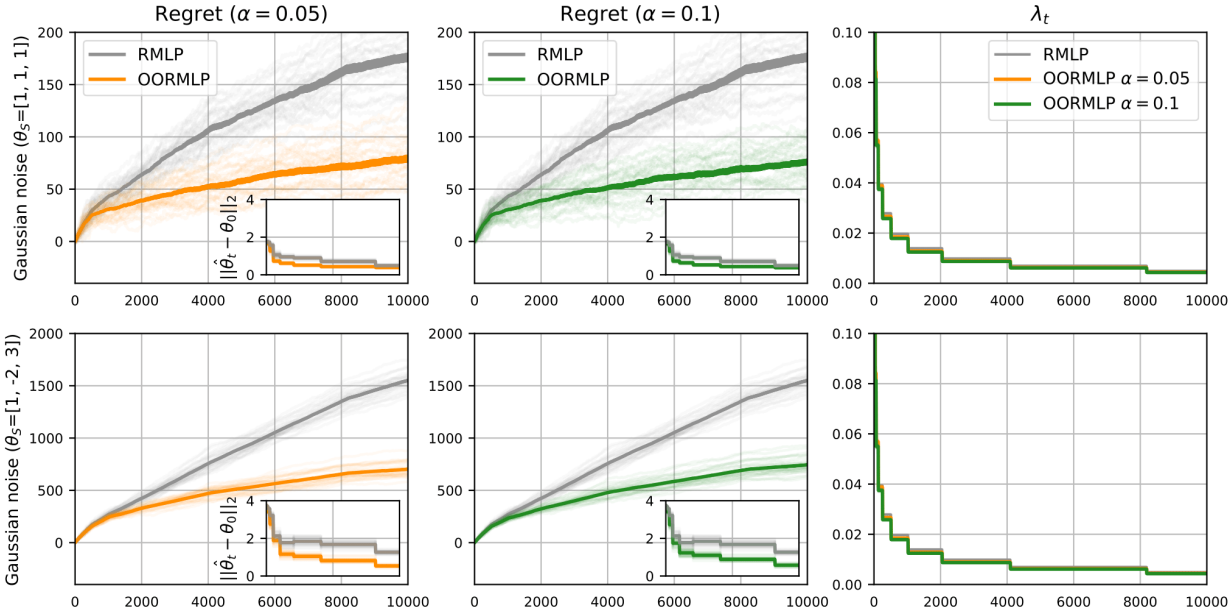

Figure 3: Comparison between `RMLP` and `OORMLP` when $d = 1000$. $\eta_t \sim N(0, 1)$. We use θ_S to denote the values in the support (the only non-zero entries) of θ_0. **First row:** $\theta_S = (1, 1, 1)$. **Second row:** $\theta_S = (1, -2, 3)$. Details for subfigures are the same as in Figure 2.

For high-dimensional experiments, we set $d = 1000$ and use the Gaussian noise setting ($\eta_t \sim N(0, 1)$) again. We consider two settings of the true demand parameter: $\theta_0 = (1, 1, 1, 0, 0, \cdots, 0)$ and $\theta_0 = (1, -2, 3, 0, 0, \cdots, 0)$. We use $c_\lambda = 0.001$ and set $W = 10$. Here to save computation resources, in both this high-dimensional setting and the real data setting below, we update the estimation of `OORMLP` only at $t = 2^k, k \in \mathbb{N}$ as in `RMLP`. Figure 3 shows the results of $t \in [0, 10000]$ over 32 replicates. `OORMLP` performs better than `RMLP` even with the same number of estimation updates. This regret reduction mainly comes from the larger sample size used by `OORMLP`. As mentioned in Remark 5, `RMLP` used a doubling trick to

apply batch-type concentration results, while our result for `OORMLP` is based on a martingale concentration. Therefore, `RMLP` only updates its estimate using the most recent batch while `OORMLP` updates its estimate using all historical information. This means the sample size used by `OORMLP` is twice larger than that used in `RMLP`, and hence `OORMLP` is more sample efficient.

6.2 Real data analysis on auto loan applications

We demonstrate the efficiency of `OORMLP` in setting personalized lending rates for an online auto loan company in the United States. Personalization of prices in the lending industry is widely used and well-accepted. Our experiments are based on a real-life data set *CPRM-12-001: On-Line Auto Lending* provided by the Center for Pricing and Revenue Management at Columbia University. This database contains data on all 208,805 auto loan applications received by a major online lender in the United States between July 2002 and November 2004. The data collection contains the date on which prospective borrowers submitted an application, the sort of loan they requested (term and amount), and some personal information. Additionally, the data collection includes whether the online lender authorized the application, the annual percentage rate (APR) given and whether a contract was executed. In this context, clients' demand responses are binary, indicating whether or not a loan was agreed upon. This dataset was studied in many dynamic pricing literatures, e.g., Phillips et al. (2015), Ban and Keskin (2021), Bastani et al. (2021).

A summary of the data set (with descriptive statistics on the demand and available features) is shown in the Table 3 in Ban and Keskin (2021). The column "apply" is the binary demand indicator for eventual contract and is the response variable with value in $\{0, 1\}$ for the market value model. There are 18 feature variables, both discrete categorical (e.g., type of financing, type of car, customer state) and continuous (e.g., FICO score, customer rate, competitor's rate). We prepossess the categorical variable to dummy variables and normalize the continuous variable to values with mean 0 and maximum absolute value 1.

This pricing problem is a special instance of the problem formulation in Section 2.2, with demand being a binary variable. In this situation, the price of a loan is determined by subtracting the loan amount from the net present value of future payments. Formally, we can calculate the price from the other variables in the dataset through $p = \text{Monthly Payment} \times \sum_{\tau=1}^{\text{Term}}(1+\text{Rate})^{-\tau} - \text{Loan Amount}$. Here, we use one thousand dollars as a basic unit for the price p. Also, note that the dimension of the variables in this dataset is $d = 71$ since we construct dummy variables from the categorical variables.

In practice, it is hard to retrieve real-time feedback from clients on any dynamic pricing strategy until the pricing policy has been implemented in the data collection system. Thus, we apply off-policy learning used in Ban and Keskin (2021) to estimate the customer choice model using $\widehat{\theta} \equiv \arg\min \mathcal{L}(\theta)$ where $\mathcal{L}(\theta)$ is defined by (8a) but across the entire dataset with the assumption η_t being i.i.d. following $N(0,1)$. This optimization problem is the same as (8b) with $\lambda_t = 0$ and $W = \infty$. We use (6) with $\theta_0 = \widehat{\theta}$ as the ground truth model for generating the response of each consumer given any price. More specifically, to generate data from this model, we sample the covariates x_t from the original dataset and η_t from $N(0,1)$, then we calculate the market value v_t and the response y_t using (2) and (6).

Similar to the simulation study in Section 6.1, we design the pricing function $g(\cdot)$ by assuming the uncertainty is standard normal ($\eta_t \sim N(0,1)$). Since the ground truth model has $\|\theta_0\|_1 = 33.68$, we use $W = 100$ as the upper bound for $\|\theta\|_1$ in the online estimation of θ for both `OORMLP` and `RMLP`. The scaling parameter is $c_\lambda = 0.00001$ and we update the estimation of $\widehat{\theta}_t$ at $t = 2^k, k \in \mathbb{N}$. We compare `OORMLP` to `RMLP` using experiment with $t \in [0, 5000]$ over 32 replicates. Figure 4 reports the results for the regret, the estimation error, and the regularization sequence used.

While both `OORMLP` and `RMLP` enjoy sublinear growth of regret, `OORMLP` obtains more accurate and stable estimation of θ and much less regret than `RMLP` at $T = 5000$ time periods across all confidence budgets under similar regularization sequence $\{\lambda_t\}_{t=1}^{5000}$. This is consistent

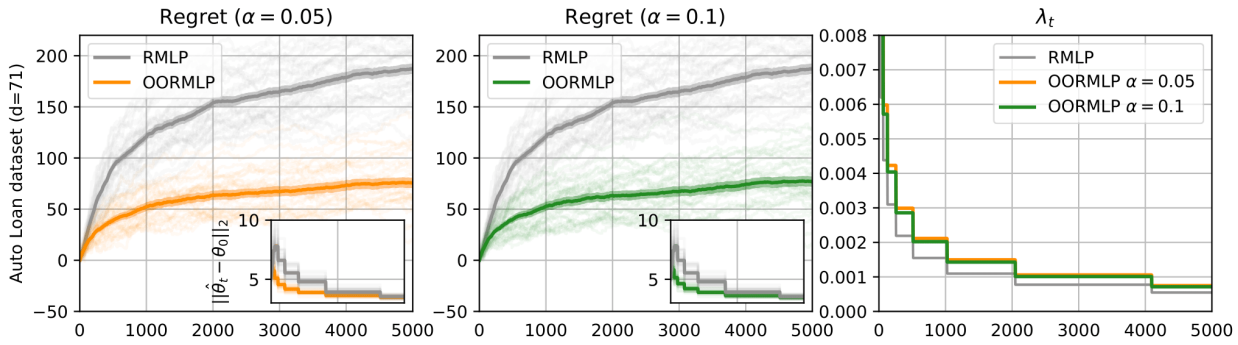

Figure 4: Comparison between `RMLP` and `OORMLP` on the On-Line Auto Lending dataset. Details for subfigures are the same as in Figure 2.

to our observation on the comparison results with synthetic data. These results support that `OORMLP` enjoys substantial further regret reduction compared to `RMLP` and supports the claimed superiority of the proposed online regularization scheme.

Acknowledgment

The authors thank the editor Professor Jane-Ling Wang, the associate editor and two anonymous reviewers for their valuable comments and suggestions which led to a much improved paper. Will Wei Sun's research was partially supported by NSF-SES grant (2217440). Guang Cheng's research was partially supported by ONR grant (N00014-22-1-2680) and NSF-SCALE MoDL grant (2134209). Any opinions, findings, and conclusions expressed in this material are those of the authors and do not reflect the views of the Office of Naval Research or the National Science Foundation. The authors report there are no competing interests to declare.

References

Backhaus, K., Becker, J., Beverungen, D., Frohs, M., Müller, O., Weddeling, M., Knackstedt, R. and Steiner, M. (2010). Enabling individualized recommendations and dynamic pricing of value-added services through willingness-to-pay data. *Electronic Markets* **20** 131–146.

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

Ban, G.-Y. and Keskin, N. B. (2021). Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science* **67** 5549–5568.

Bastani, H. and Bayati, M. (2020). Online decision making with high-dimensional covariates. *Operations Research* **68** 276–294.

Bastani, H., Simchi-Levi, D. and Zhu, R. (2021). Meta dynamic pricing: Transfer learning across experiments. *Management Science* .

Bobkov, S., Madiman, M. et al. (2011). Concentration of the information in data with log-concave distributions. *The Annals of Probability* **39** 1528–1543.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

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

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

Chen, H., Lu, W. and Song, R. (2021a). Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association* **116** 240–255.

Chen, H., Lu, W. and Song, R. (2021b). Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association* **116** 708–719.

Chen, X., Miao, S. and Wang, Y. (2021c). Differential privacy in personalized pricing with nonparametric demand models. *Available at SSRN 3919807* .

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

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

Chen, X. and Wang, Y. (2020). Uncertainty quantification for demand prediction in contextual dynamic pricing. *arXiv preprint arXiv:2003.07017* .

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

Chichignoud, M., Lederer, J. and Wainwright, M. J. (2016). A practical scheme and fast algorithm to tune the lasso with optimality guarantees. *The Journal of Machine Learning Research* **17** 8162–8181.

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

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

Fan, J., Guo, Y. and Yu, M. (2021). Policy optimization using semiparametric models for dynamic pricing. *Available at SSRN 3922825* .

Fan, J., Guo, Y. and Yu, M. (2022). Policy optimization using semiparametric models for dynamic pricing. *Journal of the American Statistical Association* 1–29.

Hazan, E. (2019). Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207* .

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2020). Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys* **17** 257–317.

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. *arXiv preprint arXiv:1901.01030* .

Johari, R., Koomen, P., Pekelis, L. and Walsh, D. (2021). Always valid inference: Continuous monitoring of a/b tests. *Operations Research* .

Johari, R., Pekelis, L. and Walsh, D. J. (2015). Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922* .

Joskow, P. L. and Wolfram, C. D. (2012). Dynamic pricing of electricity. *American Economic Review* **102** 381–385.

Kale, S., Karnin, Z., Liang, T. and Pál, D. (2017). Adaptive feature selection: Computationally efficient online sparse linear regression under rip. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

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

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.* IEEE.

Langford, J., Li, L. and Zhang, T. (2009). Sparse online learning via truncated gradient. *Journal of Machine Learning Research* **10** 777–801.

Lederer, J. (2022). Tuning-parameter calibration. In *Fundamentals of High-Dimensional Statistics*. Springer, 109–137.

Lederer, J. and Vogt, M. (2021). Estimating the lasso's effective noise. *Journal of Machine Learning Research* **22** 1–32.

Lederer, J., Yu, L., Gaynanova, I. et al. (2019). Oracle inequalities for high-dimensional prediction. *Bernoulli* **25** 1225–1255.

Leme, R. P. and Schneider, J. (2022). Contextual search via intrinsic volumes. *SIAM Journal on Computing* **51** 1096–1125.

Liu, P.-Y., Wang, C.-H. and Tsai, H.-H. (2022). Non-stationary dynamic pricing via actor-critic information-directed pricing. *arXiv preprint arXiv:2208.09372* .

Luo, Y., Sun, W. W. et al. (2021). Distribution-free contextual dynamic pricing. *arXiv preprint arXiv:2109.07340* .

Maillard, O.-A. (2019). *Mathematics of Statistical Sequential Decision Making*. Habilitation à diriger des recherches, Université de Lille, Sciences et Technologies.
URL https://hal.archives-ouvertes.fr/tel-02162189

Mao, J., Leme, R. and Schneider, J. (2018). Contextual pricing for lipschitz buyers. *Advances in Neural Information Processing Systems* **31**.

Mueller, J., Syrgkanis, V. and Taddy, M. (2019). Low-rank bandit methods for high-dimensional dynamic pricing. In *Advances in Neural Information Processing Systems*.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of operations research* **6** 58–73.

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

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B. et al. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science* **27** 538–557.

Oh, M.-h., Iyengar, G. and Zeevi, A. (2020). Sparsity-agnostic lasso bandit. *arXiv preprint arXiv:2007.08477* .

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 659–677.

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* .

Rothschild, M. (1974). A two-armed bandit theory of market pricing. *Journal of Economic Theory* **9** 185–202.

Saumard, A. and Wellner, J. A. (2014). Log-concavity and strong log-concavity: a review. *Statistics surveys* **8** 45.

Shah, V., Blanchet, J. and Johari, R. (2019). Semi-parametric dynamic contextual pricing. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

Shalev-Shwartz, S. and Singer, Y. (2007). A primal-dual perspective of online learning algorithms. *Machine Learning* **69** 115–142.

Shalev-Shwartz, S. and Tewari, A. (2011). Stochastic methods for l1-regularized loss minimization. *Journal of Machine Learning Research* **12** 1865–1892.

Su, W. J. and Zhu, Y. (2018). Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876* .

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267–288.

Usmanova, I., Krause, A. and Kamgarpour, M. (2019). Safe convex learning under uncertain constraints. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR.

Ville, J. (1939). Etude critique de la notion de collectif. *Bull. Amer. Math. Soc* **45** 824.

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

Wang, C.-H. and Cheng, G. (2020). Online batch decision-making with high-dimensional covariates. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Wang, C.-H. and Li, W. (2022). Always valid risk monitoring for online matrix completion. *arXiv preprint arXiv:2211.10363* .

Wang, C.-H., Yu, Y., Hao, B. and Cheng, G. (2020a). Residual bootstrap exploration for bandit algorithms. *arXiv preprint arXiv:2002.08436* .

Wang, J.-K., Lu, C.-J. and Lin, S.-D. (2019). Online linear optimization with sparsity constraints. In *Algorithmic Learning Theory*.

Wang, L., Peng, B., Bradic, J., Li, R. and Wu, Y. (2020b). A tuning-free robust and efficient approach to high-dimensional regression. *Journal of the American Statistical Association* **115** 1700–1714.

Wang, Y., Chen, X., Chang, X. and Ge, D. (2021). Uncertainty quantification for demand prediction in contextual dynamic pricing. *Production and Operations Management* **30** 1703–1717.

Wellner, J. (2012). Log-concave distributions: definitions, properties, and consequences. *Presentation, University of Paris-Diderot* .

Wu, S., Wang, C.-H., Li, Y. and Cheng, G. (2022). Residual bootstrap exploration for stochastic linear bandit. In *Uncertainty in Artificial Intelligence*. PMLR.

Wu, Y. and Wang, L. (2020). A survey of tuning parameter selection for high-dimensional regression. *Annual Review of Statistics and Its Application* **7** 209–226.

Xu, J. and Wang, Y.-X. (2021). Logarithmic regret in feature-based dynamic pricing. *Advances in Neural Information Processing Systems* **34** 13898–13910.

Zhao, S., Zhou, E., Sabharwal, A. and Ermon, S. (2016). Adaptive concentration inequalities for sequential decision problems. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*.