Minimax Optimality in Contextual Dynamic Pricing with General Valuation Models

Xueping Gong[1], Wei You[2], and Jiheng Zhang[3]

[1]School of Management, Xiamen University, xgongah@xmu.edu.cn
[2]Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, weiyou@ust.hk
[3]Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, jiheng@ust.hk

August 14, 2025

Abstract

We study contextual dynamic pricing, where a decision maker posts personalized prices based on observable contexts and receives binary purchase feedback indicating whether the customer's valuation exceeds the price. Each valuation is modeled as an unknown latent function of the context, corrupted by independent and identically distributed market noise from an unknown distribution. Relying only on Lipschitz continuity of the noise distribution and bounded valuations, we propose a minimax-optimal algorithm. To accommodate the unknown distribution, our method discretizes the relevant noise range to form a finite set of candidate prices, then applies layered data partitioning to obtain confidence bounds substantially tighter than those derived via the elliptical-potential lemma. A key advantage is that estimation bias in the valuation function cancels when comparing upper confidence bounds, eliminating the need to know the Lipschitz constant. The framework extends beyond linear models to general function classes through offline regression oracles. Our regret analysis depends solely on the oracle's estimation error, typically governed by the statistical complexity of the class. These techniques yield a regret upper bound matching the minimax lower bound up to logarithmic factors. Furthermore, we refine these guarantees under additional structures—e.g., linear valuation models, second-order smoothness, sparsity, and known noise distribution or observable valuations—and compare our bounds and assumptions with prior dynamic-pricing methods. Finally, numerical experiments corroborate the theory and show clear improvements over benchmark methods.

1 Introduction

The dynamic pricing problem, which involves setting real-time prices for products or services, has received significant attention due to its practical applications and direct impact on revenue

maximization in industries such as entertainment, e-commerce (Lei et al. 2018), and transportation (Saharan et al. 2020). For an extensive review of the dynamic pricing literature, we recommend den Boer (2015). Recent research has increasingly focused on feature-based dynamic pricing models that leverage observable contexts to understand market value and design effective pricing strategies (Wang et al. 2014, Cesa-Bianchi et al. 2019, Chen et al. 2024, Fan et al. 2024, Wang and Liu 2025, Wang et al. 2025). These models capture product heterogeneity and enable personalized pricing.

At each decision time, the seller observes covariates that represent relevant product features and customer characteristics. These covariates determine the customer's valuation through an unknown valuation function and market noise. The customer purchases the product if their valuation exceeds the posted price. In standard settings where only binary purchase decisions are observed, the seller receives censored feedback about the customer's latent valuation. The goal is to set prices adaptively to maximize revenue while simultaneously learning the unknown valuation function and noise distribution.

Designing low regret policies in this setting is particularly challenging. The demand curve, often obscured by market noise, shifts continuously as covariates change, and its shape may not follow any specific parametric form. As a result, solving the contextual dynamic pricing problem requires accurate estimation of the demand function over a wide range of price-context combinations, which depends intimately on the smoothness of the noise distribution or on assumptions regarding the uniqueness of the optimal price (see Table 1). This is in stark contrast to non-contextual pricing or nonparametric bandit problems, where estimation around a single optimal price suffices. Without a pricing policy that is carefully tailored to these complexities, regret can be significantly higher.

Despite extensive research on the contextual dynamic pricing problem, a gap remains in developing policies that are provably optimal while relying on mild assumptions. Existing approaches often require strong assumptions about the smoothness of the valuation function or the noise distribution, which may not hold in practice. For instance, Javanmard and Nazerzadeh (2019) assume a known noise distribution, Tullii et al. (2024) require that the Lipschitz constant of the noise distribution is known, and Fan et al. (2024) require the noise distribution to be m-th differentiable.

We tackle these challenges through an episode-based explore-then-UCB framework. In each episode, our algorithm begins with an exploration phase that collects data to estimate the valuation function. The algorithm then enters a UCB phase, where we discretize the noise domain into equal-length intervals; this yields a finite-action linear bandit structure that supports robust upper confidence bounds capturing both valuation-estimation error and discretization bias. Within each episode, we control regret using a tighter concentration argument (via Azuma's inequality), which requires engineered independence. We ensure this independence through a carefully designed layered data partitioning scheme that accounts for model misspecification from estimation error. Notably, the adaptive nature of this technique eliminates the need for prior knowledge of the Lipschitz constant of the noise distribution, enhancing flexibility and practicality. Finally, by tuning the number of grid intervals to balance learning and discretization errors, the policy attains minimax-optimal regret.

1.1 Contributions

A novel algorithm with mild assumptions. We propose a minimax optimal algorithm for contextual dynamic pricing that operates under rather mild assumptions—specifically, bounded valuations and Lipschitz-continuous noise distributions. These conditions are standard in the literature (Javanmard and Nazerzadeh 2019, Chen and Gallego 2021, Choi et al. 2023, Fan et al. 2024, Luo et al. 2024). In contrast to prior work that relies on stronger assumptions such as known Lipschitz constants, known noise distributions, second-order smoothness of revenue functions, or uniqueness of the optimal price, our method dispenses with all of these, marking a substantial advancement in generality and applicability. Our algorithm achieves a regret upper bound[1] of $\tilde{\mathcal{O}}(\rho_{\mathcal{V}}^{\frac{1}{3}}(\delta)T^{\frac{2}{3}})$, where $\rho_{\mathcal{V}}(\delta)$ captures the statistical complexity of the valuation function space. This bound matches the lower bound up to logarithmic factors, closing the gap in the literature. As a by-product, we extend existing lower bounds for dynamic pricing problems to cover smoother distributions beyond mere Lipschitz continuity (see Theorem 2).

Improved regret bounds for linear valuation models. When applied to linear valuation models with d_0-dimensional covariates, our results yield $\tilde{\mathcal{O}}(d_0^{\frac{1}{3}}T^{\frac{2}{3}})$ regret as $\rho_{\mathcal{V}}(\delta) = \mathcal{O}(d_0 \ln(d_0/\delta))$. This improves significantly upon prior works, such as $\tilde{\mathcal{O}}(d_0^2 T^{\frac{2}{3}})$ (with additional second-order smoothness assumption) and $\tilde{\mathcal{O}}(d_0 T^{\frac{3}{4}})$ (without it) in Luo et al. (2024). Moreover, compared to Fan et al. (2024), who obtain $\tilde{\mathcal{O}}(d_0^{\frac{5}{7}}T^{\frac{5}{7}})$ regret upper bound, our approach achieves a lower order in T while relaxing assumptions. In Table 1, we provide a comprehensive comparison with existing methods under linear valuation models, highlighting the improvements achieved by our approach in terms of reduced regret bounds and relaxed assumptions.

Table 1: Comparison of Existing Methods for Linear Valuation Models with d_0 Dimensional Features.

Method	**Regret***	**Additional Assumptions†**
Luo et al. (2022, 2024)	$d_0 T^{\frac{3}{4}}$	N/A
	$d_0^2 T^{\frac{2}{3}}$	Second-order smoothness (Assumption 5)
	$d_0^2 T^{\frac{2}{3}} \vee d_0^{\frac{1}{2}} T^{1-\frac{\alpha}{2}}$	Availability of a classification oracle (Assumption 7)
Fan et al. (2024)	$(d_0 T)^{\frac{3}{4}}$	Uniqueness of the optimal price
	$(d_0 T)^{\frac{2m+1}{4m-1}}$	m-th differentiable, uniqueness of the optimal price
Tullii et al. (2024)	$(d_0 T)^{\frac{2}{3}}$	Known Lipschitz constant L in Assumption 3
Ours (corollary 1)	$d_0^{\frac{1}{3}} T^{\frac{2}{3}}$	N/A

Note:

* The regrets listed above omit constant factors, poly(ln(d_0)) and poly(ln(T)) terms.

† All methods require realizability (Assumption 1), boundedness of the valuation function (Assumption 2) and Lipschitz continuity (Assumption 3), which are therefore omitted in the table.

Generalization beyond linear models. Our method is flexible as it naturally extends beyond

[1]The notation $\tilde{\mathcal{O}}$ hides the constant factors and logarithmic terms in T and $\rho_{\mathcal{V}}(\delta)$.

linear models by considering general function spaces and leveraging general offline regression oracles. This enables us to handle a broad range of models, including sparse linear models and advanced frameworks such as reproducing kernel Hilbert spaces (RKHS) for the valuation function space. In contrast, many existing methods assume that the valuation function is linear (Luo et al. 2022, Fan et al. 2024, Javanmard and Nazerzadeh 2019), which limits their applicability to more complex settings. For example, Xu and Wang (2022) perform component-wise discretization of the parameter space, which is intrinsically tied to linearity assumption, and hence it is unclear how it may be extended to non-linear valuation structures. Cohen et al. (2020) employ ellipsoid-based shallow cuts to refine parameter uncertainty, so their geometric approach fundamentally relies on linearity. Recent work by Chen et al. (2024) relaxes linearity by proposing a nonparametric nearest-neighbor estimator for general valuation functions, which is a special offline regression oracle.

Improved regret bounds under additional information. Our framework is flexible and can incorporate additional structural information to further tighten regret bounds. We highlight three illustrative scenarios: (i) Under censored observations and access to a classification oracle (Assumption 7), our method improves upon the $\tilde{\mathcal{O}}(T^{\frac{2}{3}\vee(1-\frac{\alpha}{2})})$ bound in Luo et al. (2024), achieving a sharper rate of $\tilde{\mathcal{O}}(T^{\frac{3}{5}\vee(1-\frac{\alpha}{2})})$, where α characterizes the oracle's estimation accuracy. (ii) When the noise distribution is known, our algorithm matches the regret guarantees of Javanmard and Nazerzadeh (2019) (which focus on linear models) while extending to more general, potentially nonlinear function classes. (iii) In settings where full customer valuations (rather than binary purchase feedback) are available, and the revenue function satisfies a second-order smoothness condition, we attain a tighter regret bound of $\tilde{\mathcal{O}}(T^{\frac{3}{5}})$.

Layered data partitioning technique for dynamic pricing. To the best of our knowledge, this work is the first to introduce the layered data partitioning technique in the context of dynamic pricing. This method partitions the data into temporally decoupled layers, ensuring statistical independence across layers and allows for sharp confidence bounds via Azuma's inequality. Leveraging this technique, our algorithm achieves a regret of $\tilde{\mathcal{O}}(d_0^{\frac{1}{3}}T^{\frac{2}{3}})$ improving upon the previous best-known bound of $\tilde{\mathcal{O}}(d_0T^{\frac{3}{4}})$. Importantly, our method is parameter-free: it does not require prior knowledge of problem-specific constants (e.g., the Lipschitz constant), in contrast to approaches such as Tullii et al. (2024). This advantage arises from the observation that the estimation bias in the valuation function is common across arms and thus cancels when comparing upper confidence bounds, making our framework readily deployable in real-world settings.

1.2 Related Work

Dynamic pricing. Dynamic pricing is an active area of research, driven by advancements in data technology and the increasing availability of customer information. Initial research focused on non-contextual dynamic pricing (Besbes and Zeevi 2015, Cesa-Bianchi et al. 2019). For example, Wang et al. (2021) employed the UCB approach with local-bin approximations, achieving an $\tilde{\mathcal{O}}(T^{\frac{m+1}{2m+1}})$ regret for m-th smooth demand functions and establishing a matching lower bound. However, these approaches do not incorporate covariates into pricing policies.

In the domain of dynamic pricing with covariates, the linear customer valuation model has been widely adopted. Javanmard and Nazerzadeh (2019) studied this model assuming a *known* and *log-concave* noise distribution. In contrast, Golrezaei et al. (2019) considered an ambiguity set for the noise distribution, achieving a regret of $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$ compared to a robust benchmark, although their approach struggles when the ambiguity set encompasses an infinite number of distributions. Our model addresses the general case of an unknown noise distribution and establishes regret bounds by comparing against the true optimal policy instead of a robust benchmark. Chen and Gallego (2021) explored nonparametric aspects of the unknown demand function using adaptive binning of the covariate space to achieve a regret of $\tilde{\mathcal{O}}(T^{\frac{d_0+2}{d_0+4}})$. Notably, our method outperforms theirs when $d_0 \geq 2$. Under the Cox proportional hazards (PH) model, Choi et al. (2023) introduced the CoxCP algorithm, which achieved a regret of $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$. Their approach relies on the separability of the unknown linear structure and noise distribution in the PH model, making it unsuitable for our setting where these components are entangled. Xu and Wang (2022) proposed an adaptive pricing policy with $\tilde{\mathcal{O}}(T^{\frac{3}{4}})$ regret for adversarial contexts and bounded noise distributions. Our method improves this to $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$ under Lipschitz noise distributions, while also maintaining computational efficiency compared to the exponential computations required by the EXP4-based policy in Xu and Wang (2022). Recently, Wang and Chen (2025) established a minimax-optimal regret bound of $\tilde{\mathcal{O}}(T^{3/5})$ under twice-differentiability and additional structural conditions (e.g., strong unimodality of the revenue function). Their analysis focuses on a linear valuation model with unknown price elasticity and leverages active learning for parameter estimation. In contrast, our approach requires substantially weaker assumptions and provides a parameter-free algorithm.

Many papers investigate sparsity with high-dimensional covariates (Ren and Zhou 2024, Ban and Keskin 2021, Javanmard and Nazerzadeh 2019). Ren and Zhou (2024) study linear contextual bandits, and Ban and Keskin (2021) examine generalized linear demand models, proposing minimax-optimal policies for both known and unknown sparsity levels. However, sparsity has been largely under-explored in semiparametric contextual pricing. Javanmard and Nazerzadeh (2019) consider a related setting but require a known noise distribution. In contrast, our method extends naturally to sparse parameter vectors while accommodating an unknown noise distribution.

Other related studies (Luo et al. 2022, 2024, Fan et al. 2024) share similarities with our work in terms of settings, but differ in their assumptions about the smoothness of the noise distribution. For example, Luo et al. (2022) proposed an episode-based algorithm with regret bounds of $\tilde{\mathcal{O}}(d_0^2 T^{\frac{2}{3}})$ and $\tilde{\mathcal{O}}(d_0 T^{\frac{3}{4}})$ under different smoothness assumptions. Fan et al. (2024) considered m-th differentiable distributions and achieved a regret of $\tilde{\mathcal{O}}((d_0 T)^{\frac{2m+1}{4m-1}})$ using the Nadaraya-Watson kernel regression estimator. In comparison, our method achieves $\tilde{\mathcal{O}}(d_0^{\frac{1}{3}} T^{\frac{2}{3}})$ regrets when applied to linear valuation models, demonstrating minimax optimality while relying only on Lipschitz assumption.

Contextual bandits. Our pricing policy is closely related to bandit algorithms (Lattimore and Szepesvári 2020, Foster and Rakhlin 2020, Abbasi-Yadkori et al. 2011, Takemura et al. 2021, Auer 2002) that balance exploration and exploitation in decision-making. In particular, our approach connects to misspecified linear bandits (Auer 2002, Takemura et al. 2021). Unlike traditional bandit

algorithms, dynamic pricing must account for both the variance of estimation and the bias arising from parameter perturbations. We show that the unique structure of the dynamic pricing problem, when cast as a misspecified linear bandit, permits a more precise concentration bound, leading to improved regret bounds compared to the naive application of standard misspecified linear bandit algorithms. This improvement stems from the fact that the number of candidate prices is finite in each round. Our results highlight the importance of leveraging the distinctive structure of the pricing context to achieve optimal performance.

1.3 Notation and organization

Throughout the paper, we use the following notations. For any positive integer n, we denote the list $\{1, 2, \cdots, n\}$ as $[n]$. The cardinality of a set A is denoted by $|A|$. We use $\mathbb{I}_E$ to represent the indicator function of an event E. We denote by $\|\cdot\|_p$, for $1 \leq p \leq \infty$, the ℓ_p norm. Throughout the analysis, the notation $\tilde{\mathcal{O}}$ hides dependence on absolute constants and logarithmic terms.

In Section 2, we introduce the contextual dynamic pricing problem and present key assumptions of our approach. Section 3 presents the details of our algorithmic framework, with discussions of essential techniques and methodologies. Section 4 provides a theoretical analysis of the regret bounds for the proposed algorithm. Building on this framework, we discuss several special cases and extensions of our method in Section 5. In Section 6, we present numerical experiments comparing our methods with existing approaches. While the core ideas behind the proofs are sketched throughout the paper, complete and rigorous proofs are deferred to the Appendix for clarity and completeness.

2 Problem Formulation

In the contextual dynamic pricing setting, a potential customer arrives on the platform in each round $t \in [T]$, and the seller observes a covariate vector $\boldsymbol{x}_t \in \mathcal{X} \subset \mathbb{R}^{d_0}$ that captures relevant product features and customer characteristics. We assume that the covariates $\{\boldsymbol{x}_t\}$ are drawn i.i.d. from an unknown distribution supported on $\mathcal{X}$. After observing $\boldsymbol{x}_t$, the customer's valuation for the product is given by $v_t = v^*(\boldsymbol{x}_t) + \epsilon_t$, where $v^*(\boldsymbol{x}_t)$ is an unknown valuation function, and the noise terms $\{\epsilon_t\}_{t\in[T]}$ are independent and identically distributed according to an *unknown* cumulative distribution function F, have zero mean, and are independent of everything else.

If the random valuation v_t exceeds the posted price p_t, a sale occurs and the seller earns revenue p_t. Otherwise, if $v_t < p_t$, no sale is made and the revenue is zero. We denote the sale outcome by $y_t = \mathbb{I}\{v_t \geq p_t\}$, so that y_t follows a Bernoulli distribution with parameter $1 - F(p_t - v^*(\boldsymbol{x}_t))$. The revenue at time t is therefore $r_t = p_t y_t$. Thus, the triplet $(\boldsymbol{x}_t, p_t, y_t)$ encapsulates the key information observed in the pricing process at round t.

Given the covariate $\boldsymbol{x}_t$, setting a price p yields expected revenue

$$\mathsf{Rev}_t(p) = p(1 - F(p - v^*(\boldsymbol{x}_t))).$$

The optimal price maximizes this expected revenue:

$$p_t^* \in \operatorname*{argmax}_{p \geq 0} \mathsf{Rev}_t(p).$$

The regret at time t is defined as the difference between the expected revenue of the optimal price p_t^* and that of the chosen price p_t. Over a horizon of T rounds, the cumulative regret is

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \big(\mathsf{Rev}_t(p_t^*) - \mathsf{Rev}_t(p_t)\big) = \sum_{t=1}^{T} \Big[p_t^*\big(1 - F(p_t^* - v^*(\boldsymbol{x}_t))\big) - p_t\big(1 - F(p_t - v^*(\boldsymbol{x}_t))\big)\Big].$$

The goal in contextual dynamic pricing is to select a price p_t for each observed covariate $\boldsymbol{x}_t$, using historical data $\{(\boldsymbol{x}_s, p_s, y_s)\}_{s\in[t-1]}$, in order to learn the unknown valuation function v^* and noise distribution F, while minimizing the cumulative regret $\mathrm{Reg}(T)$.

We now outline the key assumptions used in this work. To begin, we make realizability and boundedness assumptions.

Assumption 1 (Realizability)**.** *The valuation function v^* belongs to a known function class $\mathcal{V}$.*

Assumption 2 (Bounded Valuation)**.** *There exist positive, finite constants B_ϵ and B such that the market noise is uniformly bounded: $|\epsilon_t| \leq B_\epsilon$ for all t, the valuation function is bounded away from the extremes: $v^*(\boldsymbol{x}) \in [B_\epsilon, B - B_\epsilon]$ for all $\boldsymbol{x} \in \mathcal{X}$.*

Assumption 2 imposes a known upper bound on customer valuations, which is a natural and practical assumption for real-world products. Since $v^*(\boldsymbol{x})$ is bounded above by $B - B_\epsilon$, the optimal price is also bounded above by a universal constant B.

Next, we impose a Lipschitz-continuity condition on the noise distribution F, a standard yet relatively mild assumption in the dynamic-pricing literature (Luo et al. 2022, 2024, Chen et al. 2024). Notably, this requirement is weaker than those adopted in many prior works. For instance, Chen et al. (2024), Fan et al. (2024), Javanmard and Nazerzadeh (2019) assume that F admits a bounded derivative, and Luo et al. (2022), Fan et al. (2024), Javanmard and Nazerzadeh (2019), Wang and Chen (2025) additionally require the optimal price to be unique. We require neither assumption in our work. Moreover, we do not impose any concavity conditions on F or on the revenue function. In Section 5, we discuss how stronger assumptions lead to improved results.

Assumption 3 (Lipschitz Continuity)**.** *The noise distribution F is Lipschitz continuous with a positive constant L, i.e., $|F(x) - F(y)| \leq L|x - y|, \forall x, y \in \mathbb{R}$.*

Assumptions 2 and 3 are satisfied by a broad range of distributions, such as uniform and truncated normal distributions.

3 The Algorithm

At each round t the seller sequentially observes the covariate $\boldsymbol{x}_t$ and sets a price p_t based on the history $\{\boldsymbol{x}_1, p_1, y_1, \ldots, \boldsymbol{x}_{t-1}, p_{t-1}, y_{t-1}, \boldsymbol{x}_t\}$, which inherently requires balancing exploration

(gathering informative samples to improve estimates of the unknown v^* and F) with exploitation (leveraging current estimates to maximize immediate revenue). In this section, we propose a distribution-free pricing policy (Algorithm 1) for contextual dynamic pricing with unknown valuation function v^* and noise distribution F, without requiring restrictive assumptions about these functions.

Our algorithm tackles the exploration-exploitation trade-off by decoupling the learning components: it allocates controlled exploration to collect informative price-outcome pairs for estimating v^*, then applies confidence-bound mechanisms to optimize revenue based on these estimates. Specifically, the algorithm operates in episodes, each consisting of an exploration phase, an estimation phase, and a UCB phase (see Algorithm 2). These three phases must be carefully balanced with respect to the time horizon T in order to achieve minimax-optimal regret. When T is unknown, we employ the standard doubling trick, partitioning the time horizon into exponentially growing episodes of length $\ell_k = 2^{k-1}$. The schematic of the algorithm for a single episode is shown in Figure 1.

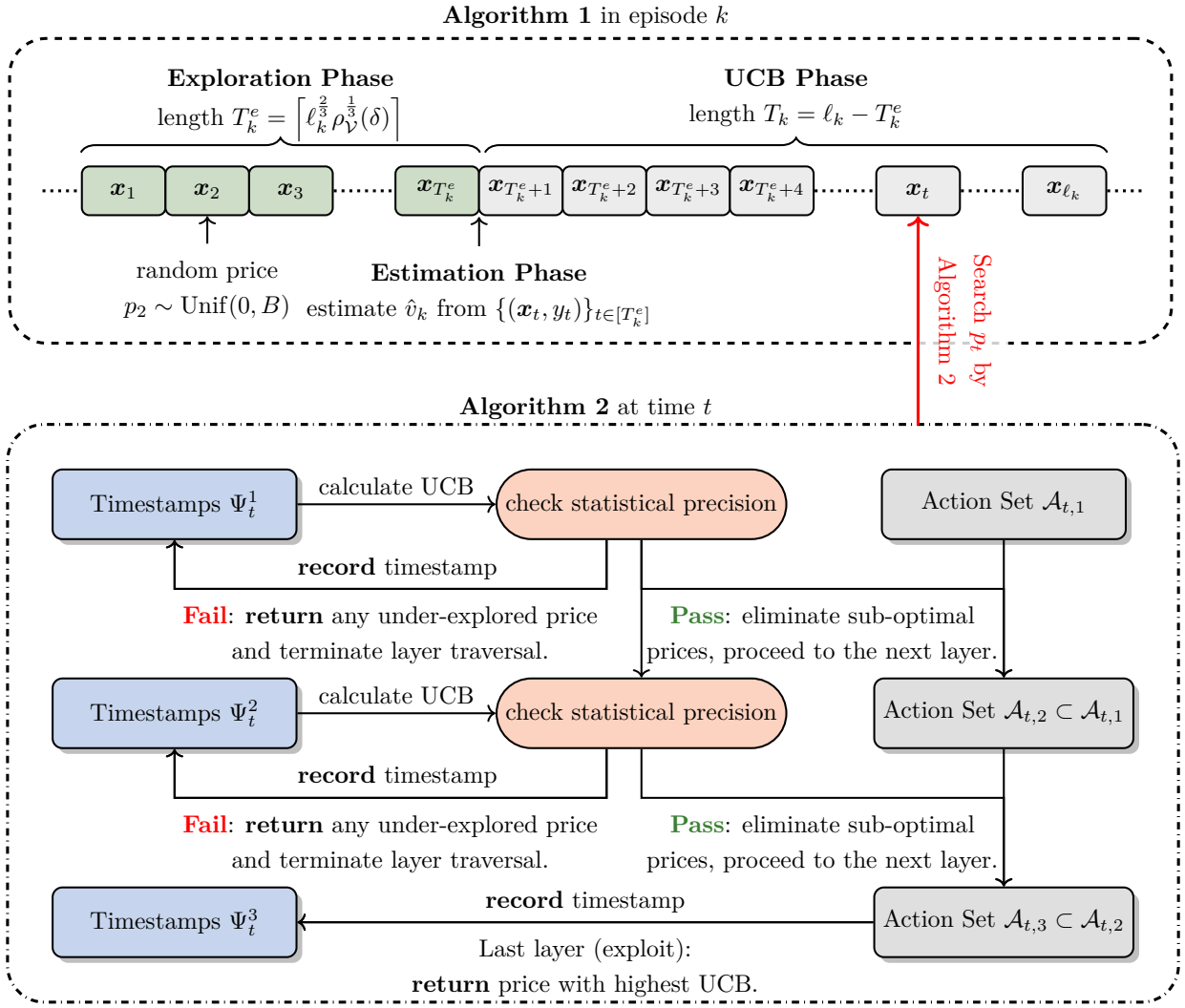

Figure 1: Schematic diagram of Algorithm 1 (dashed box) of episode k with length $\ell_k = 2^{k-1}$, and Algorithm 2 at time t (dash-dotted box) with $S_k = \left\lceil \frac{1}{2} \log_2(T_k) \right\rceil$ layers ($S_k = 3$ for illustration purpose).

Overview of Algorithm Design. Algorithm 1 leverages a novel combination of techniques to address the challenges of dynamic pricing under uncertainty.

1. **Exploration and Estimation Phases.** Estimating the valuation function v^* is inherently entangled with both the pricing policy and the unknown noise distribution F. To disentangle these elements, we adopt the approach of Javanmard and Nazerzadeh (2019) by introducing a dedicated *exploration phase* of predetermined length. During this phase, we sample prices uniformly at random from a bounded set of candidate values, ensuring thorough coverage of the covariate-price space. This design yields a clean regression structure that separates the task of estimating v^* from the influence of F, enabling an offline regression oracle to recover v^* accurately in the subsequent *estimation phase*.

2. **UCB-Phase (Algorithm 2).** Our core innovations are realized here:

   - **Sharper concentration bounds.** Under only a Lipschitz-continuity assumption on F, we discretize the noise distribution and cast the problem as a perturbed linear bandit (Luo et al. 2024). Unlike prior works, by capping the number of candidate actions per episode, we replace the conventional elliptical potential lemma with a tighter component-wise concentration bound based on Azuma's inequality.
   - **Modified layered data partitioning.** The use of Azuma's inequality is made possible by enforcing statistical independence through a layered data partitioning (LDP) technique (Auer 2002, Li et al. 2019), which partitions samples into disjoint subsets so that all confidence intervals are constructed from independent data. Whereas classical LDP assumes exact knowledge of the reward function, we develop a novel variant that accommodates the regression oracle's estimation error in v^*, preserving the required independence structure while controlling error propagation in the valuation estimates.

Collectively, these innovations yield a minimax-optimal regret bound under only mild Lipschitz-continuity assumptions. In what follows, we present a detailed exposition of the algorithm's design.

3.1 Decouple the Estimation of the Valuation Function v^*

The seller collects binary feedback y_t, which depends jointly on the posted price p_t, the valuation $v^*(\boldsymbol{x}_t)$, and the unknown noise distribution F. This coupling complicates direct estimation of v^*, as traditional methods would require simultaneous identification of F.

To decouple these components, we adopt a strategy inspired by Javanmard and Nazerzadeh (2019). Specifically, during the Exploration Phase of each episode (Step 6 of Algorithm 1), the seller posts prices sampled uniformly from $[0, B]$. This leads to the following relationship:

$$\mathbb{E}[By_t \mid \boldsymbol{x}_t] = B\mathbb{E}\big[\mathbb{E}[\mathbb{I}\{p_t \le v^*(\boldsymbol{x}_t) + \epsilon_t\} \,|\boldsymbol{x}_t, \epsilon_t] \mid \boldsymbol{x}_t\big] = B\mathbb{E}\left[\frac{v^*(\boldsymbol{x}_t) + \epsilon_t}{B}\Big|\boldsymbol{x}_t\right] = v^*(\boldsymbol{x}_t),$$

where the noise term ϵ_t vanishes due to its zero-mean property. This enables consistent estimation of v^* through offline regression (Step 9 of Algorithm 1) on exploration-phase data, independent of F. For linear valuation models, standard least squares suffices.

More generally, we rely on an offline regression oracle that satisfies Assumption 4. This abstraction separates the statistical estimation of the valuation function v^* from the exploration-exploitation trade-off in pricing.

Assumption 4 (Offline Regression Oracle)**.** *Under Assumption 1, let $\{(\boldsymbol{x}_t, y_t)\}_{t\in[n]}$ be i.i.d. samples from a fixed but unknown distribution, satisfying $\mathbb{E}[By_t \mid \boldsymbol{x}_t] = v^*(\boldsymbol{x}_t)$. Given these samples and any confidence level $\delta > 0$, an offline regression oracle returns a predictor $\hat{v} \in \mathcal{V}$ such that*

$$\|\hat{v} - v^*\|_\infty \leq \sqrt{\rho_{\mathcal{V}}(\delta)/n} \quad \text{with probability at least } 1-\delta.$$

This assumption quantifies the estimation accuracy inherent to statistical learning. The term $\rho_{\mathcal{V}}(\delta)$ captures the intrinsic complexity of learning the function class $\mathcal{V}$ as the confidence parameter δ decreases. Under the realizability (Assumption 1), the quantity $\sqrt{\rho_{\mathcal{V}}(\delta)/n}$ bounds the ℓ_∞ error between v^* and its estimator $\hat{v}$. Deriving sharp bounds on this error, and designing efficient algorithms that achieve them, are fundamental objectives in statistical learning. We provide examples of appropriate oracle constructions in Appendix A.

To ensure minimax optimality in terms of T, we set the length of the exploration phase in the episode k as $T_k^e \triangleq \lceil \ell_k^{\frac{2}{3}} \rho_{\mathcal{V}}^{\frac{1}{3}}(\delta) \rceil$, balancing estimation precision with the episode duration ℓ_k. Early episodes with $k \leq k^* \triangleq \lceil \log_2(\rho_{\mathcal{V}}(\delta)) \rceil$ use pure uniform pricing, in which case we cap T_k^e by ℓ_k.

Algorithm 1 Distribution-Free Dynamic Pricing Algorithm with Offline Regression Oracle

Input: price bound B and statistical complexity $\rho_{\mathcal{V}}(\delta)$

for round $t = 1, 2, \cdots, 2^{k^*-1}$ **do** ▷ Warm up
  Observe context $\boldsymbol{x}_t$, set a price $p_t \sim \mathrm{Unif}(0, B)$, and observe the feedback y_t
end for
for episode $k = k^*, k^*+1, \cdots$ **do**
  Set the length of the k-th episode as $\ell_k = 2^{k-1}$ and its exploration phase as $T_k^e = \lceil \ell_k^{\frac{2}{3}} \rho_{\mathcal{V}}^{\frac{1}{3}}(\delta) \rceil$
  for round $t = 2^{k-1}+1, \cdots, 2^{k-1}+T_k^e$ **do** ▷ Exploration Phase
    Observe context $\boldsymbol{x}_t$, set a price $p_t \sim \mathrm{Unif}(0, B)$, and observe the feedback y_t
  end for
  Call **Offline Regression Oracle** on $\{(\boldsymbol{x}_t, y_t)\}_{t=2^{k-1}+1}^{2^{k-1}+T_k^e}$ to get $\hat{v}_k$ ▷ Estimation Phase
  Set the length $T_k = \ell_k - T_k^e$ and the discretization number $N_k = \lceil T_k^{\frac{1}{3}} / \ln^{\frac{1}{3}}(T_k/\delta) \rceil$
  for $t = 2^{k-1}+T_k^e+1, \cdots, 2^k$ **do** ▷ UCB Phase
    Apply **UCB-LDP** (Algorithm 2) to incoming contexts $\boldsymbol{x}_t$ with the estimator $\hat{v}_k$, the discretization number N_k, the length T_k, the bound B and the confidence parameter δ
  end for
end for

Remark 1. *In Algorithm 1, the valuation function v^* is estimated using only the samples from the current episode k. While it is possible to leverage all prior exploration data, doing so does not improve the final regret bound, as each exploration phase provides a sample size of $\mathcal{O}(\ell_k^{\frac{2}{3}}\rho_{\mathcal{V}}^{\frac{1}{3}}(\delta))$, so the sample size of the current episode dominates. To simplify the algorithm's presentation, we use only the current episode's data to estimate $\hat{v}_k$.*

3.2 Discretization for Noise Distribution F

For notational simplicity and clarity, we omit the episode subscript k whenever it is unambiguous. Specifically, quantities originating from Algorithm 1 retain the subscript k (e.g., $\hat{v}_k$) to indicate episode affiliation. For auxiliary quantities introduced in Algorithm 2 and used in derivations or proofs within a single episode, we drop the subscript for brevity; the relevant episode is clear from context, so no ambiguity arises.

The regression estimate $\hat{v}_k$ is used to guide the UCB phase of each episode, which balances F-learning with revenue maximization (see Algorithm 2). We restrict F-learning to the interval $[-\|\hat{v}_k\|_\infty, B+\|\hat{v}_k\|_\infty]$ by discretizing it into N_k subintervals with midpoints $\{m_j\}_{j\in[N_k]}$. Consequently, at round t, the candidate price set is $\{m_j + \hat{v}_k(\boldsymbol{x}_t)\}_{j=1}^{N_k}$. This discretization is crucial for the UCB phase, as it induces a linear bandit structure. Specifically, at each round t we define the parameter vector

$$\xi_t = (1 - F(m_1 + \hat{v}_k(\boldsymbol{x}_t) - v^*(\boldsymbol{x}_t)), \cdots, 1 - F(m_{N_k} + \hat{v}_k(\boldsymbol{x}_t) - v^*(\boldsymbol{x}_t)))\,.$$

We represent arm j by the vector $a_j \in \mathbb{R}^{N_k}$ with $p_j := m_j + \hat{v}_k(\boldsymbol{x}_t)$ at its j-th component and zero at all other components. Posting price p_j at time t (i.e., pulling arm j) then yields expected revenue $\xi_t^\top a_j$. This corresponds to a perturbed linear bandit with nominal parameter $\xi^* = (1 - F(m_1), 1 - F(m_2), \ldots, 1 - F(m_{N_k}))$. Although each ξ_t deviates from ξ^*, Lipschitz continuity (Assumption 3) together with the oracle's error bound (Assumption 4) guarantee that these deviations remain small.

Remark 2 (Connection with the Literature)**.** *Prior work by Chen et al. (2024) and Fan et al. (2024) estimate the full CDF F using a Nadaraya-Watson kernel estimator. In contrast, our method targets only a finite set of F-values, sidestepping the complexity of nonparametrically estimating F (or its density F') over the entire domain. Focusing on discrete points allows us to obtain tighter concentration bounds at each location, which in turn yields stronger regret guarantees. Meanwhile, Luo et al. (2024) also formulate the problem as a perturbed linear bandit but relies on the elliptical potential lemma to derive its UCB, leading to suboptimal regret. Importantly, our algorithm guarantees that episode k involves at most N_k actions per round. In this finite-arm setting, the linear contextual-bandit literature (e.g., Auer 2002, Li et al. 2019) shows that one can bypass the elliptical potential lemma. These works instead employ a layered data partitioning scheme, which yields stronger regret guarantees under linear bandit models.*

Remark 3 (Challenges in Applying Layered Data Partitioning)**.** *It is challenging to attain minimax-optimality in our dynamic-pricing framework due to the following reasons.*

1. ***Parameter Normalization.*** *The standard LDP procedure rescales the parameter ξ^* by its ℓ_2 norm. However, since $\|\xi^*\|_2 = \mathcal{O}(\sqrt{N_k})$, this normalization would introduce an extra $\sqrt{N_k}$ factor into the regret bound. Because N_k typically grows polynomially in T, such a term leads to suboptimal regret scaling. We address this by exploiting the fact that each action vector has only one non-zero entry and satisfies $\left|(m_j + \hat{v}_k(\boldsymbol{x}_t))(1 - F(m_j))\right| \leq B$. Consequently, rather than normalizing ξ^* in ℓ_2, we derive UCBs using the ℓ_∞ estimation error. By enforcing statistical independence through a layered data partitioning scheme (Auer 2002, Chu et al. 2011), we can apply a tighter Azuma's inequality to obtain the required concentration bounds.*

2. ***Parameter Perturbation.*** *A second complication arises from perturbations of the nominal parameter ξ^*, which introduce misspecification into the linear-bandit model. Concretely, if at time t we observe covariate $\boldsymbol{x}_t$ and post price $p_t = m_j + \hat{v}_k(\boldsymbol{x}_t)$, then*

$$\mathbb{E}[y_t \mid \boldsymbol{x}_t, p_t] = 1 - F(m_j + \hat{v}_k(\boldsymbol{x}_t) - v^*(\boldsymbol{x}_t)).$$

This is a perturbation of the ground truth parameter $\xi_j^* = 1 - F(m_j)$. By Lipschitz continuity (Assumption 3) and the oracle's error guarantee (Assumption 4), the perturbation scales with the accuracy of the estimator. When $\hat{v}_k$ is exact, the problem reduces to a standard linear bandit, and classical LDP methods (Auer 2002, Li et al. 2019) apply directly. Larger estimation errors, however, amplify misspecification and can degrade regret (see Theorem 1), motivating our refined LDP scheme that explicitly handles model misspecification and error propagation.

3.3 Layered Data Partitioning

We now present the UCB-LDP algorithm (Algorithm 2), which directly addresses the challenges identified in Remark 3. Fix an episode k, LDP organizes past data into layers $s \in [S_k]$, where $S_k = \lceil \frac{1}{2} \log_2(T_k) \rceil$ and $T_k = \ell_k - T_k^e = 2^{k-1} - \lceil \ell_k^{\frac{2}{3}} \rho_{\mathcal{V}}^{\frac{1}{3}}(\delta) \rceil$ is the length of the UCB phase. At each time t, each layer s maintains its own dataset Ψ_t^s, containing rounds prior to time t that made their pricing decisions at layer s. The central idea of LDP is the partitioning of historical data $[t-1]$ into *disjoint layers* $\{\Psi_t^s : s \in [S_k]\}$. Since each data point belongs to exactly one layer, the confidence intervals for layer s depend only on Ψ_t^s, eliminating cross-layer dependencies. Moreover, each layer corresponds to a distinct level of statistical precision—lower layers admit wider confidence bounds. This enforced independence justifies the use of Azuma's inequality for tight concentration bounds.

We now elaborate on the details. Fix a time t, the data of each round belongs to exactly one *stopping layer* s_t, determined alongside the pricing decision p_t. The algorithm starts at layer $s = 1$ and examines successive layers until a price is selected; see Figure 1 (bottom panel) for an illustration of the algorithm's flow at time t with three layers. For each price $p_j = m_j + \hat{v}_k(\boldsymbol{x}_t)$, let $\Psi_t^s(j) \triangleq \{\tau \in \Psi_t^s \mid m_j + \hat{v}_k(\boldsymbol{x}_\tau) = p_\tau\} \subset \Psi_t^s$ be the set of previous rounds in layer s that used p_j. At layer s, we estimate ξ_j^* via the sample mean over samples from layer s only:

$$w_{t,s}^j = \frac{1}{|\Psi_t^s(j)|} \sum_{\tau \in \Psi_t^s(j)} y_\tau.$$

Algorithm 2 Upper Confidence Bound with Layered Data Partitioning (**UCB-LDP**)

Input: the estimator $\hat{v}$, the discretization number N, the length T, the price bound B and the confidence parameter δ

Divide the F-learning interval $[-\|\hat{v}\|_\infty, B + \|\hat{v}\|_\infty]$ into N equal-length intervals with their midpoints denoted as $\{m_j\}_{j\in[N]}$
Set max number of layers as $S = \left\lceil \frac{1}{2}\log_2(T) \right\rceil$, and initialize $\Psi_1^s = \emptyset$ for $s \in [S]$
for $t = 1, 2, \cdots, T$ **do**
    Observe the context $\boldsymbol{x}_t$
    Initialize current layer index $s = 1$ and action set $\mathcal{A}_{t,1} = \{j \in [N] \mid m_j + \hat{v}(\boldsymbol{x}_t) \in (0, B)\}$
    while index j_t is not found **do**
        For all $j \in [N]$, compute confidence radius $r_{t,s}^j$ defined in (1) and $\mathrm{UCB}_{t,s}^j$ defined in (2)
        if $s = S$ **then** ▷ Exploitation
            Choose the index $j_t = \mathrm{argmax}_{j\in\mathcal{A}_{t,s}} \mathrm{UCB}_{t,s}^j$
        else
            if $(m_j + \hat{v}(\boldsymbol{x}_t))r_{t,s}^j > B2^{-s}$ for some $j \in \mathcal{A}_{t,s}$ **then** ▷ Fail statistical precision check
                Choose an arbitrary j_t such that $(m_j + \hat{v}(\boldsymbol{x}_t))r_{t,s}^j > B2^{-s}$ ▷ Exploration
            else if $(m_j + \hat{v}(\boldsymbol{x}_t))r_{t,s}^j \le B2^{-s}$ for all $j \in \mathcal{A}_{t,s}$ **then** ▷ Pass statistical precision check
                Let $\mathcal{A}_{t,s+1} = \{j \in \mathcal{A}_{t,s} \mid \mathrm{UCB}_{t,s}^j \ge \max_{j'\in\mathcal{A}_{t,s}} \mathrm{UCB}_{t,s}^{j'} - B2^{1-s}\}$ ▷ Price elimination
                Advance to the next layer $s \leftarrow s + 1$
            end if
        end if
    end while
    Set the stopping layer $s_t = s$ and price $p_t = m_{j_t} + \hat{v}(\boldsymbol{x}_t)$
    Update $\Psi_{t+1}^{s_t} = \Psi_t^{s_t} \cup \{t\}$ and keep $\Psi_{t+1}^{\sigma} = \Psi_t^{\sigma}$ for $\sigma \neq s_t$
end for

The corresponding confidence radius is given by

$$r_{t,s}^j \triangleq \min\left\{\sqrt{\frac{2\ln(2S_kN_kT_k/\delta)}{|\Psi_t^s(j)|}}, 1\right\}. \tag{1}$$

We consequently compute the UCB for the revenue of price p_j at round t and layer s as

$$\mathrm{UCB}_{t,s}^j \triangleq (m_j + \hat{v}(\boldsymbol{x}_t))(w_{t,s}^j + r_{t,s}^j). \tag{2}$$

If $\Psi_t^s(j)$ is empty, indicating no previous round has selected price p_j, we set its UCB to infinity and define $w_{t,s}^j = 0$ by convention.

While traversing the layers, we maintain a candidate-action set $\mathcal{A}_{t,s}$ initialized by $\mathcal{A}_{t,1} = \{j \in [N_k] \mid m_j + \hat{v}_k(\boldsymbol{x}_t) \in (0, B)\}$ that includes all discretized prices in $(0, B)$. As we move through layers, the set of candidate actions $\mathcal{A}_{t,s}$ may shrink: $\mathcal{A}_{t,1} \supseteq \cdots \supseteq \mathcal{A}_{t,S_k}$. The algorithm's progression

through layers is guided by the level of revenue uncertainty, determined by $(m_j + \hat{v}_k(\boldsymbol{x}_t))r^j_{t,s}$, for currently active actions in $\mathcal{A}_{t,s}$:

(a) **Exploitation**: The layer traversal terminates unconditionally at the final layer S_k, in which case we must have $(m_j + \hat{v}_k(\boldsymbol{x}_t))r^j_{t,s} \le 2B/\sqrt{T_k}$ for all $j \in \mathcal{A}_{t,S_k}$. This implies that the revenue uncertainty is small for *all* remaining prices. Consequently, the algorithm selects the price p_j with the highest UCB and set the stopping layer $s_t = S_k$.

(b) **Price elimination**: For any layer $s < S_k$, if *all* prices indexed in $\mathcal{A}_{t,s}$ satisfy $(m_j+\hat{v}_k(\boldsymbol{x}_t))r^j_{t,s} \le B \cdot 2^{-s}$, we eliminate those whose UCB falls short of the maximum by at least $B2^{1-s}$, and proceed to the next layer with the remaining prices.

(c) **Exploration**: For any layer $s < S_k$, if there exists $j \in \mathcal{A}_{t,s}$ such that $(m_j+\hat{v}_k(\boldsymbol{x}_t))r^j_{t,s} > B\cdot 2^{-s}$, the revenue uncertainty is substantial. In this case, the algorithm selects any such price p_j for further exploration, halts the layer traversal, and sets the stopping layer $s_t = s$.

Because the stopping layer s_t at round t depends only on $\{\Psi^s_t\}_{s\le s_t}$, the statistical correlations across layers are decoupled. This property is formally analyzed by Li et al. (2019), Auer (2002) for linear bandits. The following lemma provides a key result on the UCB property of the algorithm.

Lemma 1. *Fix an episode k. For any round t in the episode k, and any layer s at round t, with probability at least $1 - \delta/(S_kN_kT_k)$, each $j \in [N_k]$ satisfies*

$$|\xi^*_j - w^j_{t,s}| \le r^j_{t,s} + L\eta^j_{t,s}, \quad \textit{where} \quad \eta^j_{t,s} = \frac{1}{|\Psi^s_t(j)|}\sum_{\tau\in\Psi^s_t(j)} |\hat{v}_k(\boldsymbol{x}_\tau) - v^*(\boldsymbol{x}_\tau)|.$$

Based on Lemma 1, we construct a high-probability event

$$\Gamma_k = \left\{|\xi^*_j - w^j_{t,s}| \le r^j_{t,s} + L\eta^j_{t,s}, \forall t, \forall s, \forall j\right\}. \tag{3}$$

We clearly see that $\eta^j_{t,s} \le \|\hat{v}_k - v^*\|_\infty$ for any t, s, j. Combining with a union bound implies that

$$\mathbb{P}(\overline{\Gamma}_k) \le \frac{\delta}{S_kN_kT_k} \times N_k \times T_k \times S_k = \delta.$$

The UCB property offers an intuitive decomposition: the first term captures the variance from random feedback y_t given $\boldsymbol{x}_t$ and p_t, and the second term represents the bias from estimating $\hat{v}_k$. As episodes progress, more data accumulate, reducing both variance and bias, so that $w^j_{t,s}$ converges to ξ^*_j. Crucially, our construction implies that the resulting bounds do not explicitly depend on the (unknown) Lipschitz constant for F. This follows from the fact that comparing UCB values relative to each other cancels out the bias term, which is common across all indices. Hence, our approach sidesteps the need for explicit knowledge of Lipschitz parameters, further simplifying implementation.

4 Regret Analysis

In this section, we analyze the regret of our proposed Algorithm 1. We first analyze a single episode of Algorithm 1 and then extend the analysis to the entire horizon.

4.1 Upper Bounds

The estimator $\hat{v}_k$ obtained during the estimation phase plays a crucial role in guiding the UCB procedure. A lower estimation error in $\hat{v}_k$ is expected to yield lower regret during the UCB phase. However, achieving such accuracy requires a longer exploration phase, which itself incurs regret. Therefore, in addition to the "inner" balance between revenue maximization and F-learning within the UCB phase, there is also an "outer" balance between the regret incurred during the exploration phase and that incurred during the UCB phase, both relating to learning v^*.

For each round t in a fixed episode k, we denote the best price from the discretized set as

$$\tilde{p}_t^* \triangleq m_{j_t^*} + \hat{v}_k(\boldsymbol{x}_t), \quad \text{where} \quad j_t^* = \operatorname*{argmax}_{j \in [N_k]} \mathsf{Rev}_t\left(m_j + \hat{v}_k(\boldsymbol{x}_t)\right).$$

In our regret analysis, we decompose the per-round regret into two parts: the learning and the discretization regret. In particular, the regret at round t can be written as

$$\mathsf{Rev}_t(p_t^*) - \mathsf{Rev}_t(p_t) = \underbrace{\mathsf{Rev}_t(\tilde{p}_t^*) - \mathsf{Rev}_t(p_t)}_{\text{learning regret } \mathcal{R}_t^1} + \underbrace{\mathsf{Rev}_t(p_t^*) - \mathsf{Rev}_t(\tilde{p}_t^*)}_{\text{discretization regret } \mathcal{R}_t^2}.$$

The term $\mathcal{R}_t^1$ is called the *learning regret*; it measures the loss incurred by uncertainty about F and by misspecification from using $\hat{v}_k$ in the discretized price set. The term $\mathcal{R}_t^2$ is called the *discretization regret*; it quantifies the loss due to discretizing the continuous price decision.

4.1.1 Learning Regret

We first analyze the total learning regret, denoted as $\sum_{t=1}^T \mathcal{R}_t^1$, from the inner UCB algorithm while keeping the exploration and estimation phases fixed.

Define the best among the remaining prices in $\mathcal{A}_{t,s}$ at layer s as

$$\tilde{p}_{t,s}^* \triangleq m_{j_{t,s}^*} + \hat{v}_k(\boldsymbol{x}_t), \quad \text{where} \quad j_{t,s}^* = \operatorname*{argmax}_{j \in \mathcal{A}_{t,s}} \mathsf{Rev}_t(m_j + \hat{v}_k(\boldsymbol{x}_t)).$$

If the estimation of $\hat{v}_k$ is exact, our algorithm ensures $j_{t,s}^* = j_t^*$ under the high probability event Γ_k in (3), and hence classical results in Auer (2002) can be applied, because the optimal action will not be eliminated when Γ_k holds.

However, the estimation error in $\hat{v}_k$ may result in the undesired elimination of the best discretized price for each layer s, leading to a revenue gap that propagates through the layers. Specifically, we want to bound the revenue difference between $\tilde{p}_{t,s}^*$ and $\tilde{p}_t^*$. Fortunately, this gap can be controlled with careful analysis. When advancing to a new layer, the revenue difference between the prices indexed by $j_{t,s}^*$ and $j_{t,s+1}^*$ is bounded by a constant multiple of the misspecification error, which can be upper bounded by $BL\|\hat{v}_k - v^*\|_\infty$. For any layer $s \in [s_t - 1]$, such error propagation occurs $s-1$ times, and we can prove that the revenue gap is upper bounded by $4BL(s-1)\|\hat{v}_k - v^*\|_\infty$, as stated in Lemma 2.

Lemma 2. *For each round t in episode k and each layer $s \in [s_t]$, conditional on event Γ_k, we have*

$$\mathsf{Rev}_t(\tilde{p}_t^*) - \mathsf{Rev}_t(\tilde{p}_{t,s}^*) \le 4BL(s-1)\|\hat{v}_k - v^*\|_\infty.$$

Now, we can effectively control the regret within layer s using $j^*_{t,s}$ as a benchmark. The statistical precision check (Step 13 of Algorithm 2) ensures that the confidence radius decreases exponentially with the layer index s, for all prices with indices in $\mathcal{A}_{t,s}$. The additional bias term $(4BL\|\hat{v}_k - v^*\|_\infty)$ in Lemma 3 arises from the parameter perturbation.

Lemma 3. *For each round t in episode k and each layer $2 \le s \le s_t$, conditional on event Γ_k, we have*

$$\mathsf{Rev}_t(\tilde{p}^*_{t,s}) - \mathsf{Rev}_t(p_t) \le 8B \cdot 2^{-s} + 4BL\|\hat{v}_k - v^*\|_\infty.$$

Combining Lemma 2 and Lemma 3, we immediately obtain Lemma 4, which is an upper bound on the learning regrets for prices in layer s. This bound decomposes into variance $8B \cdot 2^{-s}$ and bias $4BLs\|\hat{v}_k - v^*\|_\infty$. As we increase the value of s, the variance decreases exponentially, while the bias only increases linearly. Therefore, having a larger value of s is mostly[2] advantageous for online learning.

Lemma 4. *For each round t in episode k and each layer $2 \le s \le s_t$, conditional on event Γ_k, we have*

$$\mathsf{Rev}_t(\tilde{p}^*_t) - \mathsf{Rev}_t(p_t) \le 8B \cdot 2^{-s} + 4BLs\|\hat{v}_k - v^*\|_\infty.$$

To obtain an upper bound on the cumulative regret, we decompose the time horizon into episodes and, within each episode, into layers. For any round t and any layer s, Lemma 4 bounds the regret incurred in that round by a function of the layer index s. To bound the cumulative regret over all rounds in Ψ^s_t, we analyze the cardinality of Ψ^s_t. Specifically, we use the inequality $|\Psi^s_t| \le |\Psi^s_{T_k+1}|$ and bound the latter using Lemma 11. Therefore, the overall cumulative learning regret of the inner UCB algorithm is bounded by the sum of the regret contributions across all layers and all episodes. Formally, we provide a bound on the learning regret of the inner UCB algorithm.

Lemma 5. *Under Assumptions 1, 2 and 3, the learning regret $\sum_{t=1}^T \mathcal{R}^1_t$ is bounded by*

$$\sum_{k=1}^{\lceil \log_2 T \rceil} \Big[16B\sqrt{2N_kT_k \ln(2S_kT_kN_k/\delta) \ln T_k} + 9BL\|\hat{v}_k - v^*\|_\infty T_k \ln T_k + 4BT_k^{\frac{1}{2}} + 64BN_k \ln(2S_kT_kN_k/\delta) \Big] \quad \text{with probability at least } 1 - \lceil \log_2 T \rceil \delta.$$

The regret bound in Lemma 5 shows that higher estimation accuracy for both F and v^* leads to lower regret in the UCB phase. The first term in the bound captures the effect of uncertainty in estimating the discretized F-values. The second term in Lemma 5 reflects the estimation error of $\hat{v}_k$, i.e., the misspecification error in the linear bandit model. The last two terms are of order $\mathcal{O}(\sqrt{T})$ and are dominated by the first two terms.

Remark 4 (Comparison with Luo et al. 2022)**.** *Luo et al. (2022) introduced a perturbed linear bandit formulation for dynamic pricing and established an expected learning regret bound of $\mathcal{O}(N_k\sqrt{T_k \ln T_k})$. In comparison, we improve this bound to $\mathcal{O}(\sqrt{N_kT_k \ln T_k \ln N_k})$ by leveraging the ℓ_∞-norm to bound*

[2]Although larger s tightens variance, the traversal may stop early when precision checks fail, triggering exploration.

the revenue. Specifically, the revenue at round t takes the form $\xi_t^\top a_j$, and we may use either the ℓ_2 or the ℓ_∞ norm to bound it:

$$\xi_t^\top a_j \le \|\xi_t\|_2 \|a_j\|_2 \le B \|\xi_t\|_2 \quad \text{or} \quad \xi_t^\top a_j \le \|\xi_t\|_\infty \|a_j\|_1 \le B \|\xi_t\|_\infty .$$

The ℓ_∞-based bound takes advantage of the fact that each action vector a_j has exactly one non-zero component, thus avoiding an extra $\sqrt{N_k}$ factor incurred when using the ℓ_2 norm, as in Luo et al. (2022). However, this improvement is nontrivial: switching from the ℓ_2 to the ℓ_∞ norm precludes the use of the elliptical potential lemma, which is commonly used in linear bandit analysis. Instead, we develop concentration bounds based on Azuma's inequality. A key insight that underlies our method is that the number of candidate prices becomes finite after discretization. This allows us to construct an ℓ_∞-based UCB for each price individually. If the price set were infinite and the revenue function lacked additional structure, applying an ℓ_∞-based UCB would not be feasible.

4.1.2 Discretization Regret

We now analyze the discretization regret, denoted by $\sum_{t=1}^T \mathcal{R}_t^2$, which arises due to the discretization of the continuous price space.

Intuitively, increasing the number of price points reduces the discretization regret. A finer price grid allows the best discrete price to more closely approximate the optimal price in the continuous space. This intuition is formalized in Lemma 6, which shows that the discretization regret is inversely proportional to the number of candidate prices N_k and proportional to the number of rounds T_k.

Lemma 6. *Under Assumptions 1 and 2, the discretization regret $\sum_{t=1}^T \mathcal{R}_t^2$ is upper bounded by*

$$\sum_{t=1}^{T} \mathcal{R}_t^2 \le \sum_{k=1}^{\lceil \log_2 T \rceil} \frac{3BT_k}{N_k}.$$

4.1.3 Regret Upper Bound

We are now ready to present the overall regret upper bound of Algorithm 1. To minimize the total regret, we must optimally balance the learning regret and the discretization regret. Increasing the discretization parameter N_k reduces the discretization regret by yielding a finer approximation of the continuous price space. However, as indicated by Lemma 5, it also increases the learning regret due to the expanded set of candidate prices and the associated complexity of the search process. To achieve the optimal trade-off, we set $N_k = \tilde{\mathcal{O}}(T_k^{1/3})$, which leads to an overall regret upper bound of $\tilde{\mathcal{O}}(T^{2/3})$.

Theorem 1. *Suppose $0 < \delta < 1/(2\lceil \log_2 T \rceil)$. Under Assumptions 1, 2 and 3, the regret of Algorithm 1 satisfies*

$$\mathrm{Reg}(T) = \tilde{\mathcal{O}}\Big(\rho_{\mathcal{V}}^{\frac{1}{3}}(\delta) T^{\frac{2}{3}}\Big) \quad \text{with probability at least } 1 - 2\delta\lceil \log_2 T \rceil.$$

We remark that our regret bound depends on $\mathcal{V}$ only through the estimation error parameter $\rho_{\mathcal{V}}(\delta)$ from the offline regression oracle tailored to the finite function space; see Assumption 4. In Section 5, we provide extensions to incorporate general function spaces and discuss corresponding regret upper bound, where we also compare our results with existing works under the linear valuation model.

Our analysis of the overall regret, derived from Lemma 5 and Theorem 1, identifies four distinct sources of error: (i) the regret associated with exploration for estimating v^*, (ii) the regret incurred in learning the distribution F, (iii) the estimation error of the valuation function $\hat{v}$, and (iv) the discretization error. As summarized in Table 2, these components scale differently: T_k^e for collecting samples to estimate v^*, $\sqrt{N_k T_k}$ for learning F, $\|\hat{v}_k - v^*\|_\infty T_k$ for estimation error, and T_k/N_k for baseline discretization error. Crucially, Theorem 2 establishes the minimax optimality of our approach.

Table 2: Regret Contributions and Orders for Episode k.

Term	**Regret Order**	**Note**
Regret regarding sample collection	$\mathcal{O}(T_k^e)$	Linear in the exploration length
Regret of learning F	$\tilde{\mathcal{O}}(\sqrt{N_k T_k})$	Lemma 5
Estimation error of $\hat{v}_k$	$\tilde{\mathcal{O}}(\|\hat{v}_k - v^*\|_\infty T_k)$	Lemma 5
Discretization error	$\mathcal{O}(T_k/N_k)$	Lemma 6

4.2 Lower Bounds

Xu and Wang (2022) establish a regret lower bound of $\Omega(T^{\frac{2}{3}})$ for the non-contextual pricing problem with a Lipschitz continuous noise distribution F. Combined with the $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$ regret upper bound in Theorem 1, the minimax optimality of Algorithm 1 up to logarithmic terms is established.

We extend the lower bound for the non-contextual pricing problem under an additional assumption that the noise distribution F is m-th differentiable. Several important special cases are worth highlighting. When $m = 1$, Lipschitz continuity holds for F, and we recover the $\Omega(T^{\frac{2}{3}})$ lower bound in Xu and Wang (2022). The case $m = 2$ implies Lipschitz continuity and second-order smoothness in F, which aligns with the $\Omega(T^{\frac{3}{5}})$ regret lower bound in Luo et al. (2022). Moreover, since our constructed hard instances satisfy the requirements in Wang et al. (2021), our results also imply an $\Omega(T^{\frac{m+1}{2m+1}})$ lower bound for a general m-th differentiable demand function.

Theorem 2. *Consider a non-contextual pricing problem where the market noise is independently and identically generated from an unknown distribution. Let the distribution satisfy the following conditions:*

1. *$F(\cdot)$ is nondecreasing, right-continuous, takes values in $[0, 1]$, and m times differentiable on a bounded interval $[c_1, c_2]$.*

2. *The revenue function $\mathsf{Rev}(x) = x(1 - F(x))$ has a unique maximizer within the interval $[c_1, c_2]$.*

Then, no policy can achieve an $\mathcal{O}(T^{\frac{m+1}{2m+1}-\zeta})$ regret bound for any $\zeta > 0$, where T represents the number of pricing rounds.

To establish our lower bounds, we follow the standard roadmap for lower-bound proofs in continuum-armed bandits (Kleinberg 2004, Wang et al. 2021, Luo et al. 2022, Xu and Wang 2022). Notably, Wang et al. (2021) construct an m-times differentiable *demand function* to derive a lower bound. However, their constructed demand function is not monotone so their construction cannot directly convert to a valid cumulative distribution function. To address this limitation, we explicitly construct a valid distribution that satisfies the conditions stated in Theorem 2.

4.2.1 Sketch of Construction

Our construction builds on an infinitely differentiable base-case bump function $B(x)$ supported on $[0,1]$. This base-case bump function is then rescaled to an arbitrary interval $[a,b] \subset [0,1]$ as $B_{[a,b]}(x) = B(\frac{x-a}{b-a})$.

We now describe the remaining steps of the construction. We first construct a sequence of nested intervals $[a_k, b_k]$ with widths $w_k = 3^{-k!}$ for $k \geq 0$, such that the intersection of these intervals converges to a single point x^*. To start, we set $[a_0, b_0] = [0,1]$ with width $w_0 = 1$. For each $k \geq 1$, let $w_k = 3^{-k!}$. We divide the middle third $[a_{k-1} + w_{k-1}/3, b_{k-1} - w_{k-1}/3]$ into $Q_k = w_{k-1}/(3w_k)$ equal subintervals of length w_k. For each k, we then have Q_k possible choices of $[a_k, b_k]$. By construction the intersection $\bigcap_{k=0}^{\infty}[a_k, b_k]$ is a single point. For each choice of the sequence $\{[a_k, b_k]\}_{k=0}^{\infty}$, we can define

$$f(x) = f(x; \boldsymbol{a}, \boldsymbol{b}, c_f, m) = c_f \sum_{k=0}^{\infty} w_k^m B_{[a_k,b_k]}(x),$$

where $c_f > 0$ is chosen small enough so that $\|f'\|_\infty < 1$. Each term $w_k^m B_{[a_k,b_k]}$ is C^∞ and vanishes (with all derivatives up to order m) at the endpoints of $[a_k, b_k]$, and the rapid decay $w_k \to 0$ ensures the series converges in C^m. We refer to $f(x)$ as the *bump tower*. To normalize the range of $f(x)$ to $[0,1]$, we define the rescaled function $g(x) = \frac{f(x)}{f(x)+1}$. Using this, we define the revenue function on the interval $[b,1]$ by $\mathsf{Rev}(x) = b + (1-b)g(\frac{x-b}{1-b})$, and extend it linearly to the full pricing domain $[0, 1+b]$. The parameter $b \in (0,1)$ is selected to ensure that the corresponding cumulative distribution function $F(x) = 1 - \mathsf{Rev}(x)/x$ is nondecreasing; see (10) for its explicit form. We visualize the key functions in Figure 2 for the choice $(m, c_f) = (2, 0.05)$.

Finally, to establish the minimax lower bound, we show that any policy will frequently fail to identify the precise location of the peak of certain revenue functions. This results in an unavoidable accumulation of regret due to the policy's inability to fully exploit the potential revenue, which is deliberately induced by the careful structure of the constructed distribution functions. The full proof of Theorem 2 is provided in Appendix B.3.

Remark 5. *The Lipschitz continuity assumption is crucial as it enables us to control the misspecification error in the regret analysis. Without the Lipschitz condition, Xu and Wang (2022) proposed an algorithm with a regret bound of $\tilde{\mathcal{O}}(T^{\frac{3}{4}})$. In their work, they discretize the linear parameter space and distribution space to form the policy space, reducing the original problem into an EXP4*

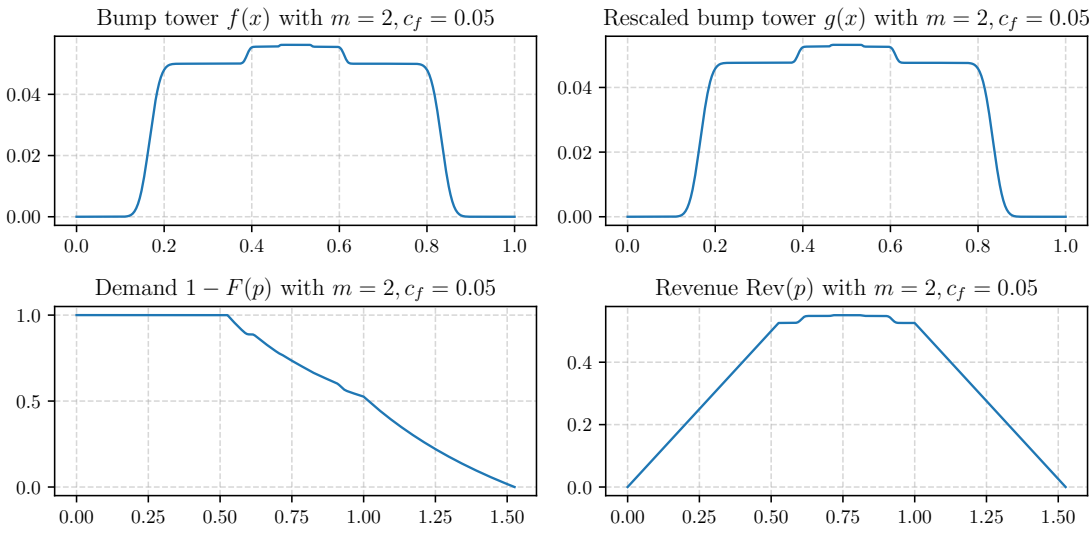

Figure 2: Construction of the Revenue Function.

agent. Although they discretize the distribution in a similar way to us, they still consider finding a good distribution within the given set, resulting in a policy space of size $\mathcal{O}(2^{N_k})$. In contrast, for setting a suitable price, we only need to evaluate at N_k points. This observation effectively reduces the search space from $\mathcal{O}(2^{N_k})$ to $\mathcal{O}(N_k)$, leading to polynomial and exponential decrease in regret and time complexity, respectively. However, it remains unclear whether their bound is tight without the Lipschitz continuity assumption. Establishing the tightness of the regret bound in this setting is an open question. Moreover, the dependence of the lower bound on the order of the statistical complexity $\rho_{\mathcal{V}}(\delta)$ remains unclear, presenting another promising direction for future research.

Remark 6. *Fan et al. (2024) propose a dynamic pricing policy for m-th differentiable F. While their work is insightful, their regret upper bound scales as $\tilde{\mathcal{O}}(T^{\frac{2m+1}{4m-1}})$, which, as shown in Theorem 2, leave room for improvement. Similarly, Chen et al. (2024) study m-th differentiable F under a general valuation model and achieve regret upper bounds of $\tilde{\mathcal{O}}(T^{\frac{2m+1}{4m-1} \vee \frac{d_0+4}{d_0+8}})$. Notably, a gap persists between these upper bounds and our derived lower bound. This gap arises for two possible reasons. First, their policy might not achieve optimality, as evidenced by the $\tilde{\mathcal{O}}(T^{\frac{3}{4}})$ regret upper bounds they incur when F is only Lipschitz, whereas a $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$ regret upper bound is attainable in such cases. Second, the inherent uncertainty in valuation models could increase the problem's complexity, potentially leading to a larger lower bound than what we derived. As a result, designing a minimax optimal policy for smoother distributions remains an open problem.*

Remark 7. *To intuitively understand the coefficients of parameters in the regret lower bound, consider a simplified scenario with a d_0-dimensional linear valuation model. Suppose we design d_0 distinct challenging sub-problems, each corresponding to a unique context. Under a uniform context distribution, each sub-problem appears approximately T/d_0 times over the horizon T. For each sub-problem, the regret lower bound scales as $\Omega(d_0^{-\frac{2}{3}} T^{\frac{2}{3}})$, which is the typical regret rate of*

non-contextual dynamic pricing problems. Aggregating the regret across all d_0 sub-problems yields a total lower bound of $\Omega(d_0^{\frac{1}{3}}T^{\frac{2}{3}})$. This illustrates how the interplay between dimension and time horizon arises: the learner must resolve uncertainty across contexts while adapting to demand shifts, further complicating the task of exploration.

5 Discussions

In this section, we refine our results under additional assumptions frequently considered in the literature and compare them with existing approaches.

5.1 Linear Valuation Models

Given the general regret bound established in Theorem 1, the function class $\mathcal{V}$ can be selected flexibly to encompass a wide range of parametric models, including linear functions, reproducing kernel Hilbert spaces, and neural networks.

Achieving practicality and optimal regret in contextual dynamic pricing requires an offline regression oracle that is both computationally efficient and statistically optimal for the chosen function class $\mathcal{V}$. To showcase the flexibility of our framework, we focus on linear function classes (Fan et al. 2024, Luo et al. 2024, Xu and Wang 2022, Golrezaei et al. 2019) and derive the corresponding regret guarantees.

In the case of linear valuation models, consider the function class $\mathcal{V} = \{\boldsymbol{\theta}^\top \boldsymbol{x} : \boldsymbol{\theta} \in \mathbb{R}^{d_0}, \|\boldsymbol{\theta}\|_2 \leq 1\}$ with $\|\boldsymbol{x}\|_2 \leq 1$, where d_0 denotes the dimension of the covariates. Under standard regularity conditions (e.g., the positive definiteness of the covariate covariance matrix), Assumption 4 holds with statistical complexity $\rho_{\mathcal{V}}(\delta) = \mathcal{O}(d_0 \ln(d_0/\delta))$. This bound is tight, as confirmed by minimax theory (Mourtada 2022).

Setting $\rho_{\mathcal{V}}(\delta) = \mathcal{O}(d_0 \ln(d_0/\delta))$ and $\delta = 1/T$ in Theorem 1, we immediately obtain the following regret upper bound for the linear valuation model.

Corollary 1. *Under Assumptions 1, 2 and 3, the expected regret of Algorithm 1 under d_0-dimensional linear valuation models satisfies*

$$\mathbb{E}[\mathrm{Reg}(T)] = \tilde{\mathcal{O}}(d_0^{\frac{1}{3}}T^{\frac{2}{3}}),$$

where the expectation is taken with respect to all the randomness in the environment.

Corollary 1 yields a substantial improvement over the previously established bound of $\tilde{\mathcal{O}}(d_0 T^{\frac{3}{4}})$ reported in Luo et al. (2024) and Xu and Wang (2022). Furthermore, when paired with the lower bound in Theorem 2, our result is minimax optimal, thereby closing the gap in the existing literature.

Remark 8 (High-Dimensional Covariates with Sparsity)**.** *Our approach extends to high-dimensional settings with sparse parameter $\boldsymbol{\theta}^*$, containing only s_0 non-zero components. Oh et al. (2021) employ Lasso regression to obtain an estimator with $\rho_{\mathcal{V}}(\delta) = \mathcal{O}(s_0 \ln(d_0/\delta))$ under mild assumptions. Javanmard and Nazerzadeh (2019) achieve similar estimation error using regularized maximum likelihood estimation. Ban and Keskin (2021) also explore sparse linear demand models in dynamic*

pricing. By exploiting sparsity, our method improves the regret upper bound from $\tilde{\mathcal{O}}(d_0^{\frac{1}{3}}T^{\frac{2}{3}})$ *to* $\tilde{\mathcal{O}}(s_0^{\frac{1}{3}}T^{\frac{2}{3}})$*, which appears to be the first such result for semi-parametric contextual pricing with unknown* F *that is only Lipschitz. Using analogous arguments, our results can be further extended to function classes in reproducing kernel Hilbert spaces, as discussed in Steinwart et al. (2009), Mendelson and Neeman (2010).*

5.2 Additional Smoothness Assumption on the Revenue Functions

Intuitively, imposing additional smoothness on the expected revenue function reduces the intrinsic complexity of dynamic pricing problems. In particular, second-order smoothness near the optimal price is a common assumption in the literature (Luo et al. 2024, Chen and Gallego 2021, Luo et al. 2022). When the revenue function is twice differentiable and the optimal price lies in the interior of the pricing range $[0, B]$, a standard Taylor expansion around the maximizer motivates this assumption. Notice that such a second-order condition implicitly requires that the optimal price be unique for every covariate. We write $p^*(\boldsymbol{x})$ for the optimal price at covariate $\boldsymbol{x}$.

Assumption 5 (Second-Order Smoothness)**.** *Define the general expected revenue function associated with the noise distribution* F *as* $\mathsf{Rev}_q(p) = p\,(1 - F(p-q))$*. There exists a positive constant* C *such that for any* $\boldsymbol{x} \in \mathcal{X}$ *and* $q = v^*(\boldsymbol{x})$*,* $\mathsf{Rev}_q\left(p^*(\boldsymbol{x})\right) - \mathsf{Rev}_q(p) \leq C\left(p^*(\boldsymbol{x}) - p\right)^2$*.*

As shown in Theorem 2 and by Luo et al. (2022), the dynamic pricing problems satisfying both Assumption 3 and Assumption 5 have a regret lower bound of at least $\Omega(T^{\frac{3}{5}})$, which demonstrates that the second-order smoothness assumption is non-trivial and effectively reduces the difficulty of dynamic pricing problems.

Remark 9. *Our method achieves a regret bound of* $\tilde{\mathcal{O}}(d_0^{\frac{1}{3}}T^{\frac{2}{3}})$ *without relying on the second-order smoothness assumption and linear valuation models. In contrast, Luo et al. (2022) leverage Assumption 5 of the revenue function to attain a regret bound of* $\tilde{\mathcal{O}}(d_0T^{\frac{2}{3}})$*. Fan et al. (2024) study a related dynamic pricing problem under the assumption that the noise distribution* F *has bounded second derivatives, which implies Assumption 5, and derive a regret bound of* $\tilde{\mathcal{O}}(T^{\frac{5}{7}})$*. The higher regret in Fan et al. (2024) stems from the need to estimate the derivative* F' *in order to compute the price at each round, introducing additional estimation error. In contrast, our approach avoids this by estimating* F *over a discrete grid. Recently, Wang and Chen (2025) provide a tight bound* $\tilde{\mathcal{O}}(T^{\frac{3}{5}})$ *under additional assumptions (twice differentiability and strong unimodality), which are stronger than ours in Assumption 3 and 5.*

5.3 Estimation Errors of Valuation Models

To further reduce regret, we consider stronger smoothness assumptions on the distribution function F. A commonly adopted condition is the second-order smoothness assumption (see Assumption 5). Under this assumption, the discretization error improves from the baseline rate of $\mathcal{O}(T_k/N_k)$ (see Table 2) to $\mathcal{O}(T_k/N_k^2)$, because the difference between the optimal price p_t^* and the optimal discretized

price $\tilde{p}_t^*$ is bounded by $\mathcal{O}(1/N_k)$. Since our theoretical analysis relies solely on the estimation error guarantee of the regression oracle (Assumption 4), any refinement in the error $\|\hat{v}_k - v^*\|_\infty$, or in the regret incurred during the sample collection phase, directly yields a tighter upper bound on the total regret. This oracle-based framework thus enables broad adaptability to various function classes, while preserving minimax optimality as established in Theorem 2.

5.3.1 General Regression Oracle

We first extend Assumption 4 by allowing the error to scale differently with the number of samples n.

Assumption 6 (General Offline Regression Oracle)**.** *Under Assumption 1, let $\{(\boldsymbol{x}_t, y_t)\}_{t\in[n]}$ be i.i.d. samples from a fixed but unknown distribution, satisfying $\mathbb{E}[By_t \mid \boldsymbol{x}_t] = v^*(\boldsymbol{x}_t)$. Given any confidence level $\delta > 0$, a general offline regression oracle returns a predictor $\hat{v} \in \mathcal{V}$ such that*

$$\|\hat{v} - v^*\|_\infty \le \sqrt{\rho_{\mathcal{V}}(\delta)/n^\alpha} \quad \text{with probability at least } 1-\delta.$$

In Chen et al. (2024), the authors construct Distributional Nearest Neighbor (DNN) and two-scale Distributed Nearest Neighbor (TDNN) estimators that satisfy Assumption 6 with rates $\alpha = \frac{4}{d_0+4}$ and $\alpha = \frac{8}{d_0+8}$ in Assumption 6, respectively.

Assuming the availability of such an offline regression oracle, we extend our results under a general estimation assumption via a simple corollary. By setting $T_k^e = \rho_{\mathcal{V}}^{\frac{1}{2+\alpha}}(\delta)\ell_k^{\frac{2}{2+\alpha}}$, we immediately obtain a regret bound by Theorem 1 and Lemma 5.

Corollary 2. *Suppose $0 < \delta < 1/(2\lceil\log_2(T)\rceil)$. Under Assumptions 1, 2, 3 and 6, the regret of Algorithm 1 with $T_k^e = \rho_{\mathcal{V}}^{\frac{1}{2+\alpha}}(\delta)\ell_k^{\frac{2}{2+\alpha}}$ satisfies*

$$\mathrm{Reg}(T) = \tilde{\mathcal{O}}\big((\rho_{\mathcal{V}}^{\frac{1}{3}}(\delta)T^{\frac{2}{3}}) \vee (\rho_{\mathcal{V}}^{\frac{1}{2+\alpha}}(\delta)T^{\frac{2}{2+\alpha}})\big) \quad \text{with probability at least } 1 - 2\lceil\log_2(T)\rceil\delta.$$

Remark 10. *Whereas Chen et al. (2024) achieve only linear $\mathcal{O}(T)$ regret under the assumption of Lipschitz continuity for F, our method attains sublinear regret rates of $\tilde{\mathcal{O}}\Big(T^{\frac{2}{3}\vee\frac{d_0+4}{d_0+6}}\Big)$ and $\tilde{\mathcal{O}}\Big(T^{\frac{2}{3}\vee\frac{d_0+8}{d_0+12}}\Big)$ when we plug in their DNN and TDNN estimators, respectively.*

5.3.2 Online Classification Oracle

An alternative approach to estimating the unknown value function v^* is to use a powerful online classification oracle, leveraging the data $(\boldsymbol{x}_t, p_t, y_t)$ collected during the $(k-1)$-th episode by an adaptive algorithm. This yields an estimate $\hat{v}_k$, which can then be used for pricing decisions in the k-th episode. To motivate this approach, observe that the outcome y_t is binary and governed by the relation $\mathbb{P}(y_t = 1) = 1 - F(p_t - v^*(\boldsymbol{x}_t))$ so that

$$\mathbb{P}(y_t = 1)\begin{cases} > \frac{1}{2}, & \text{if } F^{-1}\left(\frac{1}{2}\right) + v^*(\boldsymbol{x}_t) - p_t > 0, \\ = \frac{1}{2}, & \text{if } F^{-1}\left(\frac{1}{2}\right) + v^*(\boldsymbol{x}_t) - p_t = 0, \\ < \frac{1}{2}, & \text{if } F^{-1}\left(\frac{1}{2}\right) + v^*(\boldsymbol{x}_t) - p_t < 0. \end{cases}$$

This formulation naturally suggests a binary classification task, where the goal is to infer the decision boundary defined implicitly by the value function v^*. By invoking a classification oracle, we can obtain a predictor $\hat{v} \in \mathcal{V}$ that approximates the true value function v^*, enabling informed pricing decisions in subsequent episodes. We summarize our algorithm using a classification oracle in Algorithm 3.

Assumption 7 (Online Classification Oracle)**.** *Given data $\{(\boldsymbol{x}_t, p_t, y_t)\}_{t\in[n]}$, generated by a dynamic pricing policy, satisfying $\mathbb{P}(y_t = 1) = 1 - F(p_t - v^*(\boldsymbol{x}_t))$, and any confidence level $\delta > 0$, an online classification oracle returns a predictor $\hat{v} \in \mathcal{V}$ such that*

$$\|\hat{v} - v^*\|_\infty \leq \sqrt{\rho_{\mathcal{V}}(\delta)/n^\alpha} \quad \textit{with probability at least } 1 - \delta.$$

Algorithm 3 Distribution-Free Dynamic Pricing Algorithm with Online Classification Oracle

Input: price upper bound B, statistical complexity $\rho_{\mathcal{V}}(\delta)$, confidence parameter δ

for episode $k = 1, 2, \cdots$ **do**
  Set the length of the k-th UCB phase as $T_k = 2^{k-1}$
  Call **Online Classification Oracle** on data $\{(\boldsymbol{x}_t, p_t, y_t)\}$ from the previous episode to get $\hat{v}_k$. For $k = 1$, we set $\hat{v}_k \equiv 0$. ▷ Estimation phase
  Set the discretization number $N_k = \lceil T_k^{1/5} \rceil$
  for $t = 2^{k-1} + 1, \cdots, 2^k$ **do** ▷ UCB phase
    Apply **UCB-LDP** (Algorithm 2) on the coming sequential contexts $\boldsymbol{x}_t$ with the estimator $\hat{v}_k$, the discretization number N_k, the length T_k, the bound B and the confidence parameter δ
  end for
end for

Under the Lipschitz condition alone, the regret upper bound for the contextual dynamic pricing problem is $\tilde{\mathcal{O}}(\rho_{\mathcal{V}}^{\frac{1}{3}}(\delta)T^{\frac{2}{3}})$. However, by incorporating the additional second-order smoothness assumption (see Assumption 5), we can further improve performance, achieving tighter bounds than those established in prior work such as Luo et al. (2024).

Corollary 3. *Suppose $0 < \delta < 1/(2\lceil \log_2 T \rceil)$. Under Assumptions 1, 2, 5 and 7, the regret of Algorithm 3 satisfies*

$$\mathrm{Reg}(T) = \tilde{\mathcal{O}}(T^{\frac{3}{5}} \vee \rho_{\mathcal{V}}^{\frac{1}{2}}(\delta)T^{1-\frac{\alpha}{2}}) \quad \textit{with probability at least } 1 - 2\lceil \log_2(T) \rceil \delta.$$

Remark 11. *Wang and Chen (2025) formulate the estimation of their linear valuation model v^* as a classification problem and propose an online method inspired by active learning techniques (Chen et al. 2023). Assuming that the noise distribution F is twice differentiable, they construct an estimator $\hat{v}$ satisfying $\|v^* - \hat{v}\|_2 = \mathcal{O}(\varsigma)$, while achieving cumulative regret $\widetilde{\mathcal{O}}(d_0^3/\varsigma^2)$, where ς controls the estimation error. In contrast, if a powerful online classification oracle is available, we relax the smoothness requirements on F allowing it to be merely Lipschitz continuous.*

Remark 12. *We acknowledge that Assumption 7 is rather strong, requiring a classification oracle that can guarantee a specific estimation error with input of datasets collected by adaptive algorithms. In the work of Luo et al. (2024), the authors take a similar approach: a linear classification method is applied to the adaptively collected data from the previous episode to estimate the parameters of a linear value function. They numerically show that Assumption 7 holds with $\alpha \geq 1$ under linear valuation models.*[3] *However, it remains an open question whether a suitable classification oracle can be constructed to satisfy Assumption 7 with theoretical guarantees. As noted by Luo et al. (2022), "α is indeterministic and no rigorous justification has been made." While both Luo et al. (2024) and our work assume access to a classification oracle (they use classical logistic regression), our approach improves upon the regret guarantees in Luo et al. (2024) and matches the minimax lower bound $\Omega(T^{\frac{3}{5}})$ established in Luo et al. (2022), when the function class $\mathcal{V}$ is linear.*

5.4 Explicit Knowledge of the Distribution or Valuation Function

5.4.1 Knowledge of F

When the distribution F is fully known to the seller, Javanmard and Nazerzadeh (2019) propose an algorithm that achieves a regret of $\mathcal{O}(d_0 \ln T)$ under smoothness assumptions on F. They further show that if the covariate covariance matrix is not positive definite, an alternative approach can still guarantee a regret of $\mathcal{O}(\sqrt{T \ln d_0})$. These results highlight the significant advantage of having knowledge of F, which enables sharper regret guarantees.

Building upon this framework, we extend their methodology by incorporating an offline regression oracle to develop a distribution-dependent Algorithm 4. It is worth noting that we need to slightly refine Assumption 4 for the known-F setting. Assumption 4 requires i.i.d. data with the moment condition $\mathbb{E}[By_t \mid \boldsymbol{x}_t] = v^*(\boldsymbol{x}_t)$. When F is known, pricing decisions depend only on the context, and since the contexts are i.i.d., the observations within an episode remain i.i.d.; however, the moment condition $\mathbb{E}[By_t \mid \boldsymbol{x}_t] = v^*(\boldsymbol{x}_t)$ need not hold. Accordingly, we replace Assumption 4 with a *known-F* offline regression oracle (e.g., an MLE oracle under the correctly specified Bernoulli model with link induced by F), which still guarantees $\|\hat{v} - v^*\|_\infty \leq \sqrt{\rho_{\mathcal{V}}(\delta)/n}$ with probability at least $1 - \delta$.

The regret upper bound of Algorithm 4 is summarized in the following corollary.

Corollary 4. *Assume that the noise distribution F is twice continuously differentiable, and that both F and $1 - F$ are log-concave. Under Assumptions 1 and 4, the regret of Algorithm 4 satisfies*

$$\mathrm{Reg}(T) = \mathcal{O}\big(\rho_{\mathcal{V}}(\delta) \ln T\big) \quad \textit{with probability at least } 1 - \lceil \log_2 T \rceil \delta.$$

When the differentiability and log-concavity assumptions are relaxed, we obtain the following weaker guarantee under the Lipschitz condition alone.

[3] It is worth noting that our algorithm achieves minimax optimality as long as $\alpha \geq \frac{4}{5}$, which is slightly below the numerically justified threshold of 1.

Algorithm 4 Distribution-Dependent Dynamic Pricing Algorithm

Input: price upper bound B, distribution F, confidence parameter δ

1: **for** episode $k = 1, 2, \cdots$ **do**
2: Call **Known-F Offline Regression Oracle** on data $\{(\boldsymbol{x}_t, y_t)\}$ from the previous episode to get $\hat{v}_k$. For $k = 1$, we set $\hat{v}_k \equiv 0$.
3: **for** round $t = 2^{k-1} + 1, \cdots, 2^k$ **do**
4: Observe context $\boldsymbol{x}_t$, set price $p_t = \operatorname{argmax}_p p(1 - F(p - \hat{v}_k(\boldsymbol{x}_t)))$ and observe feedback y_t
5: **end for**
6: **end for**

Corollary 5. *Under Assumptions 1, 3 and 4, the regret of Algorithm 4 satisfies*

$$\mathrm{Reg}(T) = \mathcal{O}(\sqrt{\rho_{\mathcal{V}}(\delta) T \ln T}) \quad \textit{with probability at least } 1 - \lceil \log_2 T \rceil \delta.$$

Remark 13. *Cohen et al. (2020) also examine the linear valuation model under the assumption of a known noise distribution. Unlike our setting, they assume sub-Gaussian noise rather than enforcing the Lipschitz continuity condition (Assumption 3). Their method leverages a shallow-cut technique for ellipsoidal uncertainty sets, preserving more than half of the original volume at each iteration. Using this approach, they establish a regret bound of $\mathcal{O}(d_0^{\frac{19}{6}} T^{\frac{2}{3}} \ln^{\frac{11}{6}} T)$. However, their analysis is tailored specifically to linear valuation models and does not extend naturally to more general function classes. In contrast, our framework accommodates a broader class of valuation functions by reducing the problem to offline regression over general function spaces, improving flexibility and generalizability.*

5.4.2 Knowledge of v_t

In many practical applications, sellers may directly observe customer valuations v_t, for example, through bidding mechanisms. This additional information simplifies algorithm design by removing the need for an explicit exploration phase. In this setting, we modify Algorithm 1 to use the observed tuples $(\boldsymbol{x}_t, v_t)$ instead of $(\boldsymbol{x}_t, y_t)$ when invoking the offline oracle[4]. Unfortunately, with only the Lipschitz continuity assumption (Assumption 3), the regret of the modified algorithm *does not improve* and remains $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$.

However, with additional smoothness assumption (Assumption 5), we can achieve a tighter regret bound when the seller observes v_t. The key reason is that the discretization regret in episode k improves to $\mathcal{O}\left(T_k / N_k^2\right)$, leading to tighter overall bounds. We summarize the algorithm in Algorithm 5 and establish its regret bound in corollary 6.

[4]Similar to the known-F setting, we require an offline oracle that satisfies a variant of Assumption 4 in which the moment condition is replaced by $\mathbb{E}[v_t \mid \boldsymbol{x}_t] = v^*(\boldsymbol{x}_t)$. The formal statement is given as Assumption 10 in the Appendix.

Algorithm 5 Distribution-Free Dynamic Pricing Algorithm with General Offline Regression Oracle and Observable Valuation

Input: price upper bound B, estimation parameter $\rho_{\mathcal{V}}(\delta)$, confidence parameter δ

1: **for** episode $k = 1, 2, \cdots$ **do**
2: Call **Adjusted Offline Regression Oracle** on data $\{(\boldsymbol{x}_t, v_t)\}$ from the previous episode to get $\hat{v}_k$. For $k = 1$, we set $\hat{v}_k \equiv 0$.
3: Set the length $T_k = 2^{k-1}$ and the discretization number $N_k = \lceil T_k^{\frac{1}{5}} \rceil$
4: **for** $t = 2^{k-1} + 1, \cdots, 2^k$ **do**
5: Apply **UCB-LDP** (Algorithm 2) on the coming sequential contexts $\boldsymbol{x}_t$ with the estimator $\hat{v}_k$, the discretization number N_k, the length T_k, the bound B and the confidence parameter δ
6: **end for**
7: **end for**

Corollary 6. *Suppose* $0 < \delta < 1/(2\lceil \log_2 T \rceil)$. *Under Assumptions 1, 2, 10 and 5, the regret of Algorithm 5 satisfies*

$$\mathrm{Reg}(T) = \tilde{\mathcal{O}}\big(T^{\frac{3}{5}} \vee \rho_{\mathcal{V}}^{\frac{1}{2}}(\delta) T^{\frac{1}{2}}\big) \quad \textit{with probability at least } 1 - 2\lceil \log_2(T) \rceil \delta.$$

Assuming access only to censored demand data (i.e., observing y_t rather than the full valuation v_t), Luo et al. (2022, 2024) establish a minimax regret lower bound of $\Omega(T^{\frac{3}{5}})$. Whether access to full valuation information can lead to strictly smaller lower bounds remains an open question. We note, however, that in Section 5.3.2, we show that this lower bound is attainable even when only binary purchase decisions y_t are observed, provided that a powerful classification oracle is available.

To conclude our discussion in this section, we summarize the regret results of our proposed algorithms and compare them with existing literature in Table 3.

6 Numerical Experiments

In this section, we present numerical simulations to evaluate the empirical performance of our algorithm for linear valuation models.

6.1 Comparison with Existing Methods

We consider the contextual dynamic pricing problem under randomly generated linear valuation models. The contexts are drawn from a standard Gaussian distribution in $\mathbb{R}^4$, followed by ℓ_2-normalization to a unit ball. The noise term is drawn from a Gaussian distribution $\mathcal{N}(0, 0.3)$, truncated between -1 and 1. Similarly, the parameter θ is initialized as a normalized vector, with its components drawn from a standard Gaussian distribution.

We compare our method with the algorithms proposed in Tullii et al. (2024) and Fan et al. (2024). We note that the algorithm in Tullii et al. (2024) requires knowledge of the Lipschitz constant L in Assumption 3. For consistency with their experimental setup, we adopt $L = 1$ for their method

Table 3: Summary of Regret Guarantees and Comparison with Prior Work

Algorithm	**Regret**	**Assumptions**	**Comparison with Literature**
Algorithm 1	$\tilde{\mathcal{O}}\big((\rho_{\mathcal{V}}^{\frac{1}{3}}(\delta)T^{\frac{2}{3}}) \vee (\rho_{\mathcal{V}}^{\frac{1}{2+\alpha}}(\delta)T^{\frac{2}{2+\alpha}})\big)$	Assumption 6	DNN: $\alpha = 4/(d_0+4) \Rightarrow \tilde{\mathcal{O}}(T^{\frac{2}{3}} \vee T^{\frac{d_0+4}{d_0+6}})$; TDNN: $\alpha = 8/(d_0+8) \Rightarrow \tilde{\mathcal{O}}(T^{\frac{2}{3}} \vee T^{\frac{d_0+8}{d_0+12}})$; Chen et al. (2024) do not guarantee sublinear regret when F is only Lipschitz.
Algorithm 3	$\tilde{\mathcal{O}}\big((\rho_{\mathcal{V}}^{\frac{1}{2}}(\delta)T^{1-\frac{\alpha}{2}}) \vee T^{\frac{3}{5}}\big)$	Assumption 5 Assumption 7	Improves on $\tilde{\mathcal{O}}(T^{\frac{2}{3}} \vee T^{1-\frac{\alpha}{2}})$ in Luo et al. (2024); minimax-optimal for $\alpha \geq 4/5$.
Algorithm 4	$\mathcal{O}\big(\rho_{\mathcal{V}}(\delta)\ln T\big)$	Known F F is log-concave, twice differentiable Assumption 4	For linear $\mathcal{V}$, matches Javanmard and Nazerzadeh (2019)'s $\mathcal{O}(d_0 \ln T)$; we extend to general $\mathcal{V}$.
Algorithm 4	$\tilde{\mathcal{O}}\big(\sqrt{\rho_{\mathcal{V}}(\delta)T}\big)$	Known F Assumption 4	Cohen et al. (2020) obtain $\mathcal{O}(d_0^{\frac{19}{6}}T^{\frac{2}{3}}\ln^{\frac{11}{6}}T)$ under linear model with sub-Gaussian noise; we extend beyond linear models.
Algorithm 5	$\tilde{\mathcal{O}}(T^{\frac{3}{5}} \vee \rho_{\mathcal{V}}^{\frac{1}{2}}(\delta)T^{\frac{1}{2}})$	Assumption 5 Observable v_t	Matches the lower bound $\Omega(T^{\frac{3}{5}})$ of Luo et al. (2022, 2024) for binary feedback.

Note: Common assumptions such as Assumption 1, Assumption 2, and Assumption 3 are omitted.

while the effective constant is ≈ 0.78 (computed from the truncated normal's PDF at 0). For both baselines, Tullii et al. (2024) and Fan et al. (2024), we use the public implementation released by Tullii et al. (2024). We evaluate all methods under the price bound $B = 2$ for all time horizons $T \in \{1000, 5000, 10000, 20000, 50000\}$. The results are averaged across 10 independently generated instances.

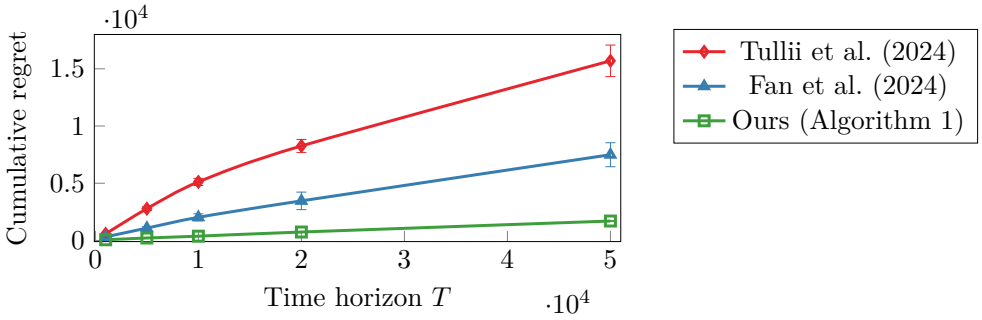

Figure 3: Comparison of the expected cumulative regret of our algorithm with Fan et al. (2024), Tullii et al. (2024). The results are averaged over 10 independent random instances, with 95% confidence intervals shown.

The average cumulative regret achieved by the three algorithms is shown in Figure 3, together with the 95% confidence intervals for the mean regret. Our algorithm consistently demonstrates superior cumulative regret (up to 80–90% lower) compared to both benchmarks across all time

horizons. Notably, the variance across 10 runs, as indicated by the confidence intervals, is significantly smaller than that of Tullii et al. (2024) and Fan et al. (2024), reflecting enhanced stability. This improvement is attributed to our adaptive layer search mechanism, which dynamically optimizes upper confidence bounds without requiring prior knowledge of the Lipschitz constant.

While Tullii et al. (2024) achieves faster runtime through Lipschitz-constant-guided UCB simplification, our method eliminates this dependency through layer-wise UCB at only 1.8–4.5 times the computation time of Tullii et al. (2024). This is particularly impactful in real-world pricing systems where Lipschitz constants are rarely known a priori. Furthermore, we significantly outperform Fan et al. (2024) in speed by avoiding their computationally intensive full distribution estimation.

6.2 Noise Distribution

In this experiment, we investigate the robustness of our algorithm's performance to different types of noise distributions. We conduct simulations using four zero-mean noise distributions: the normal distributions with variance parameter 0.5 or 1, the Cauchy distribution with scale parameter 0.1, and the uniform distribution on $[-2, 2]$.

For each of these noise distributions, we generate random context vectors of dimension $d_0 = 4$ and set the price bound $B = 4$, as in Section 6.1. The noise distributions are all truncated to the same interval of $[-3, 3]$. With the confidence parameter δ set to 0.05, we have $\rho_{\mathcal{V}}(\delta) = d_0 \ln(d_0/\delta)$. We test our algorithm over time horizons $T \in \{1000, 2000, 4000, 8000, 16000\}$. The results are reported in Figure 4, which is averaged over 10 independent replications with 95% confidence intervals presented.

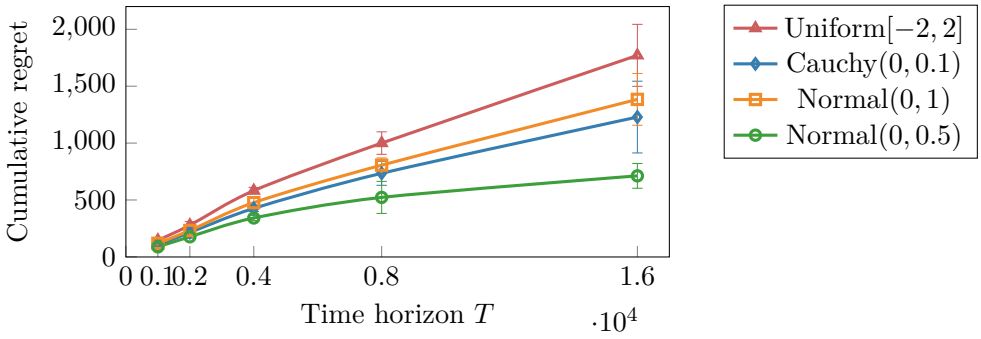

Figure 4: Comparison of the mean cumulative regret under various noise distributions. The results are averaged over 10 independent random instances, with 95% confidence intervals shown.

Figure 4 demonstrates that both the noise distribution type and variance parameter affect our algorithm's performance. We remark that our algorithm is agnostic to this distributional information, as long as the basic Lipschitz continuity Assumption 3 holds. As shown in Javanmard and Nazerzadeh (2019), the regret rate can be reduced to $\tilde{\mathcal{O}}(\sqrt{T})$ if the decision maker knows that the noise distribution belongs to the exponential family. Fan et al. (2024), Chen et al. (2024) further demonstrate that knowledge of the smoothness level of the CDF can also reduce regret. However,

the true smoothness level is typically unknown and must be determined via cross-validation. In contrast to these methods, our approach does not rely on knowing either the noise distribution family (beyond Lipschitz continuity), the CDF smoothness level, the scale parameter, or the Lipschitz constant in Assumption 3.

7 Conclusion

We present a comprehensive solution to the contextual dynamic pricing problem that combines minimax-optimality with practical applicability. Our algorithm integrates an explore-then-UCB strategy with layered data partitioning, achieving a regret upper bound of $\tilde{\mathcal{O}}(\rho_{\mathcal{V}}^{\frac{1}{3}}(\delta)T^{\frac{2}{3}})$. It improves upon existing bounds for linear valuation models and extends naturally to more general function spaces via suitable offline regression oracles, relying only on the Lipschitz continuity of the noise distribution.

Future research can pursue several directions. First, relaxing the Lipschitz continuity assumption on the noise distribution F and determining the corresponding regret lower bound would deepen our understanding of the problem's intrinsic difficulty. While Xu and Wang (2022) establish an upper bound of $\tilde{\mathcal{O}}(T^{\frac{3}{4}})$ without assuming Lipschitz continuity, it remains an open question whether the minimax lower bound in this setting is also $\Omega(T^{\frac{3}{4}})$ when the Lipschitz continuity assumption is removed.

Second, it is important to further investigate how to tighten the dependence on function-class complexity. For linear valuation models with covariates of dimension d_0, we establish a regret bound of $\tilde{\mathcal{O}}(d_0^{\frac{1}{3}}T^{\frac{2}{3}})$. Whether this $d_0^{\frac{1}{3}}$ dependence is optimal remains an open question, and resolving it would clarify how regret scales with the complexity of the underlying valuation function space.

Finally, improving regret upper bounds for smoother distributions is a promising direction. Under stronger assumptions (e.g., twice-differentiability and strong uni-modality of the revenue function) than Assumption 3, rates of order $\tilde{\mathcal{O}}(T^{3/5})$ can be achieved (Wang and Chen 2025). Further investigation in this area would clarify the impact of distribution smoothness on regret performance.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Gah-Yi Ban and N Bora Keskin. Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science*, 67(9):5549–5568, 2021.

Omar Besbes and Assaf Zeevi. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science*, 61(4):723–739, 2015.

Nicolo Cesa-Bianchi, Tommaso Cesari, and Vianney Perchet. Dynamic pricing with finitely many unknown valuations. In *Algorithmic Learning Theory*, pages 247–273. PMLR, 2019.

Elynn Chen, Xi Chen, Lan Gao, and Jiayu Li. Dynamic contextual pricing with doubly non-parametric random utility models. *arXiv preprint arXiv:2405.06866*, 2024.

Ningyuan Chen and Guillermo Gallego. Nonparametric pricing analytics with customer covariates. *Operations Research*, 69(3):974–984, 2021.

Xi Chen, Quanquan Liu, and Yining Wang. Active learning for contextual search with binary feedback. *Management Science*, 69(4):2165–2181, 2023.

Young-Geun Choi, Gi-Soo Kim, Choi Yunseo, Wooseong Cho, Myunghee Cho Paik, and Min-hwan Oh. Semi-parametric contextual pricing algorithm using cox proportional hazards model. In *International Conference on Machine Learning*, pages 5771–5786. PMLR, 2023.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Maxime C Cohen, Ilan Lobel, and Renato Paes Leme. Feature-based dynamic pricing. *Management Science*, 66(11):4921–4943, 2020.

Jianqing Fan, Yongyi Guo, and Mengxin Yu. Policy optimization using semiparametric models for dynamic pricing. *Journal of the American Statistical Association*, 119(545):552–564, 2024.

Dylan Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

Negin Golrezaei, Adel Javanmard, and Vahab Mirrokni. Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Advances in Neural Information Processing Systems*, 32, 2019.

Adel Javanmard and Hamid Nazerzadeh. Dynamic pricing in high-dimensions. *Journal of Machine Learning Research*, 20(9):1–49, 2019.

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 17, 2004.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Yanzhe Lei, Stefanus Jasin, and Amitabh Sinha. Joint dynamic pricing and order fulfillment for e-commerce retailers. *Manufacturing & service operations management*, 20(2):269–284, 2018.

Yingkai Li, Yining Wang, and Yuan Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*, pages 2173–2174. PMLR, 2019.

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

Yiyun Luo, Will Wei Sun, and Yufeng Liu. Contextual dynamic pricing with unknown noise: Explore-then-UCB strategy and improved regrets. *Advances in Neural Information Processing Systems*, 35: 37445–37457, 2022.

Yiyun Luo, Will Wei Sun, and Yufeng Liu. Distribution-free contextual dynamic pricing. *Mathematics of Operations Research*, 49(1):599–618, 2024.

Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1): 526–565, 2010.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157–2178, 2022.

Min-hwan Oh, Garud Iyengar, and Assaf Zeevi. Sparsity-agnostic Lasso bandit. In *International Conference on Machine Learning*, pages 8271–8280. PMLR, 2021.

Zhimei Ren and Zhengyuan Zhou. Dynamic batch learning in high-dimensional sparse linear contextual bandits. *Management Science*, 70(2):1315–1342, 2024.

Sandeep Saharan, Seema Bawa, and Neeraj Kumar. Dynamic pricing techniques for intelligent transportation system in smart cities: A systematic review. *Computer Communications*, 150:603–625, 2020.

Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In *COLT*, pages 79–93, 2009.

Kei Takemura, Shinji Ito, Daisuke Hatano, Hanna Sumita, Takuro Fukunaga, Naonori Kakimura, and Ken-ichi Kawarabayashi. A parameter-free algorithm for misspecified linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3367–3375. PMLR, 2021.

Matilde Tullii, Solenne Gaucher, Nadav Merlis, and Vianney Perchet. Improved algorithms for contextual dynamic pricing. In *ICML 2024 Workshop: Foundations of Reinforcement Learning and Control–Connections and Perspectives*, 2024.

Hanzhao Wang, Kalyan Talluri, and Xiaocheng Li. On dynamic pricing with covariates. *Operations Research*, 2025.

Yining Wang and Boxiao Chen. Tight regret bounds in contextual pricing with semi-parametric demand learning. *Available at SSRN 5133677*, 2025.

Yining Wang and Quanquan Liu. Estimation of high-dimensional contextual pricing models with nonparametric price confounders. *Operations Research*, 2025.

Yining Wang, Boxiao Chen, and David Simchi-Levi. Multimodal dynamic pricing. *Management Science*, 67 (10):6136–6152, 2021.

Zizhuo Wang, Shiming Deng, and Yinyu Ye. Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research*, 62(2):318–331, 2014.

Jianyu Xu and Yu-Xiang Wang. Towards agnostic feature-based dynamic pricing: Linear policies vs linear valuation with unknown noise. In *International Conference on Artificial Intelligence and Statistics*, pages 9643–9662. PMLR, 2022.

Appendix

Contents

A Examples of Offline Regression Oracle Satisfying Assumption 4

A.1 Finite-Dimensional Parameterization

We consider the case where the true function is parameterized by a finite-dimensional vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$. Specifically, we assume the true function can be expressed as:

$$v^*(\boldsymbol{x}) = g(\boldsymbol{\theta}^*, \boldsymbol{x})$$

where $g : \Theta \times \mathcal{X} \to \mathbb{R}$ is a known function. To estimate the unknown parameter $\boldsymbol{\theta}^*$, we can use a regression model that minimizes the empirical risk:

$$R_n(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - g(\boldsymbol{\theta}, \boldsymbol{x}_i))^2.$$

Then the least squares estimator is defined as $\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta} \in \Theta} R_n(\boldsymbol{\theta})$, and the estimate of v^* is given by

$$\hat{v}(x) = g(\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}).$$

For a matrix A, recall that the operator norm of A is defined as

$$\|A\|_{\mathrm{op}} = \sup_{\|\boldsymbol{u}\|_2 = 1} \|A\boldsymbol{u}\|_2.$$

Lemma 7 (Matrix Bernstein Inequality)**.** *Let $X_1, \ldots, X_n$ be independent, mean-zero, symmetric random matrices in $\mathbb{R}^{d \times d}$. Suppose that almost surely $\|X_i\|_{\mathrm{op}} \le R$, and define the variance parameter $\sigma^2 = \left\|\sum_{i=1}^n \mathbb{E}[X_i^2]\right\|_{\mathrm{op}}$. Then for all $t \ge 0$,*

$$\mathbb{P}\Big\{\Big\|\sum_{i=1}^{n} X_i\Big\|_{\mathrm{op}} \ge t\Big\} \le 2d \exp\left(-\frac{t^2}{2\sigma^2 + 2Rt/3}\right).$$

Theorem 3. *Assume the following regularity conditions hold:*

1. *$\Theta \subset \mathbb{R}^d$ is convex and compact, and $\boldsymbol{\theta}^* \in \mathrm{int}(\Theta)$.*
2. *$\sup_{\boldsymbol{\theta} \in \Theta, x \in \mathcal{X}} \|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, \boldsymbol{x})\|_2 \le L < \infty$.*
3. *$\sup_{\boldsymbol{\theta} \in \Theta, x \in \mathcal{X}} \|\nabla^2 g(\boldsymbol{\theta}, \boldsymbol{x})\|_{\mathrm{op}} \le B < \infty$.*
4. *$H(\boldsymbol{\theta}^*) = \mathbb{E}[\nabla g(\boldsymbol{\theta}^*, \boldsymbol{x}) \nabla g(\boldsymbol{\theta}^*, \boldsymbol{x})^\top] \succ \lambda_0 \boldsymbol{I}$ for $\lambda_0 > 0$.*
5. *$\|\nabla^2 g(\boldsymbol{\theta}_1, \boldsymbol{x}) - \nabla^2 g(\boldsymbol{\theta}_2, \boldsymbol{x})\|_{\mathrm{op}} \le M \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$.*
6. *$\mathcal{X} \subset \mathbb{R}^{d_0}$ is compact, and $\boldsymbol{x}_i$ are drawn i.i.d. from some distribution supported on $\mathcal{X}$.*
7. *ϵ_i are σ-sub-Gaussian with $\mathbb{E}[\epsilon_i] = 0$*

Then there exist constants $c_1, c_2 > 0$ depending on $(\lambda_0, L, B, \sigma, M, d)$ such that: for any $\delta > 0$, if $n \geq \frac{1}{c_2} \ln\left(\frac{8d}{\delta}\right)$, then with probability at least $1 - \delta$:

$$\|\hat{v} - v^*\|_\infty \leq \sqrt{\frac{d}{c_1 n} \ln\left(\frac{8d}{\delta}\right)}$$

where $c_1 = \frac{\lambda_0^2}{8\sigma^2 L^4}$ and $c_2 = \min\left(\frac{\lambda_0^2}{512L^4}, \frac{\lambda_0^2}{512\sigma^2 B^2}, \frac{c_1 \delta_0^2}{d}\right)$ with $\delta_0 = \min\left(1, \frac{\lambda_0}{16M}\right)$.

Proof. We prove the theorem through a series of probabilistic bounds.

We first prove the gradient concentration at $\boldsymbol{\theta}^*$. Define $\boldsymbol{Z}_i = \nabla g(\boldsymbol{\theta}^*, \boldsymbol{x}_i)$ and thus $\|\boldsymbol{Z}_i\|_2 \leq L$. The empirical gradient is:

$$\nabla R_n(\boldsymbol{\theta}^*) = -\frac{1}{n} \sum_{i=1}^n \epsilon_i \boldsymbol{Z}_i.$$

Since ϵ_i are σ-sub-Gaussian and independent of $\boldsymbol{x}_i$, each component of $\epsilon_i \boldsymbol{Z}_i$ is σL-sub-Gaussian. Let $\boldsymbol{S} = \sum_{i=1}^n \epsilon_i \boldsymbol{Z}_i$. Then we have:

$$\begin{aligned}
\mathbb{P}(\|\nabla R_n(\boldsymbol{\theta}^*)\|_2 \geq t\sqrt{d/n}) &= \mathbb{P}(\|\boldsymbol{S}\|_2 \geq t\sqrt{nd}) \\
&\leq \mathbb{P}\Big(\max_{1 \leq j \leq d} |S_j| \geq t\sqrt{n}\Big) \\
&\leq \sum_{j=1}^d \mathbb{P}(|S_j| \geq t\sqrt{n}) \\
&\leq \sum_{j=1}^d 2\exp\Big(-\frac{(t\sqrt{n})^2}{2\mathrm{Var}(S_j)}\Big) \\
&\leq 2d\exp\Big(-\frac{t^2 n}{2\sigma^2 n L^2}\Big) = 2d\exp\Big(-\frac{t^2}{2\sigma^2 L^2}\Big). \qquad (4)
\end{aligned}$$

Then we need to consider the Hessian matrix. Define $\delta_0 = \min\left(1, \frac{\lambda_0}{16M}\right)$. Decompose the Hessian at $\boldsymbol{\theta}^*$:

$$\nabla^2 R_n(\boldsymbol{\theta}^*) - H(\boldsymbol{\theta}^*) = \underbrace{\frac{1}{n} \sum_{i=1}^n \left[\nabla g_i \nabla g_i^\top - \mathbb{E}[\nabla g_i \nabla g_i^\top]\right]}_{T_1} - \underbrace{\frac{1}{n} \sum_{i=1}^n \epsilon_i \nabla^2 g_i}_{T_2}.$$

Since $\|\nabla g_i \nabla g_i^\top\|_{\mathrm{op}} \leq L^2$, matrix Bernstein inequality (Lemma 7) gives:

$$\mathbb{P}\left(\|T_1\|_{\mathrm{op}} \geq \frac{\lambda_0}{16}\right) \leq 2d\exp\left(-\frac{n\lambda_0^2}{512L^4}\right). \qquad (5)$$

Since $\|\epsilon_i \nabla^2 g_i\|_{\mathrm{op}} \leq |\epsilon_i| B$ and ϵ_i is σ-sub-Gaussian:

$$\mathbb{P}\left(\|T_2\|_{\mathrm{op}} \geq \frac{\lambda_0}{16}\right) \leq 2d\exp\left(-\frac{n\lambda_0^2}{512\sigma^2 B^2}\right). \qquad (6)$$

By Lipschitz continuity (Regularity Condition 5), for $\boldsymbol{\theta} \in B_{\delta_0}(\boldsymbol{\theta}^*)$:

$$\|\nabla^2 R_n(\boldsymbol{\theta}) - \nabla^2 R_n(\boldsymbol{\theta}^*)\|_{\mathrm{op}} \leq M\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq M\delta_0.$$

On the event where $\|T_1\|_{\text{op}} < \lambda_0/16$ and $\|T_2\|_{\text{op}} < \lambda_0/16$, we have

$$\|\nabla^2 R_n(\boldsymbol{\theta}) - H(\boldsymbol{\theta}^*)\|_{\text{op}} \leq M\delta_0 + \frac{\lambda_0}{8} \leq \frac{\lambda_0}{16} + \frac{\lambda_0}{8} = \frac{3\lambda_0}{16} < \frac{\lambda_0}{4}.$$

Thus, it holds that $\lambda_{\min}(\nabla^2 R_n(\boldsymbol{\theta})) \geq \lambda_0 - \lambda_0/4 > \lambda_0/2$. By union bound over (5) and (6):

$$\mathbb{P}\left(\inf_{\boldsymbol{\theta}\in B_{\delta_0}(\boldsymbol{\theta}^*)} \lambda_{\min}(\nabla^2 R_n(\boldsymbol{\theta})) \geq \lambda_0/2\right) \geq 1 - 2d\exp\left(-\frac{\lambda_0^2 n}{512L^4}\right) - 2d\exp\left(-\frac{\lambda_0^2 n}{512\sigma^2 B^2}\right). \quad (7)$$

With (4) and (7), we can now bound the parameter estimation error. By Taylor expansion and first-order optimality for empirical risk minimization:

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* = -[\nabla^2 R_n(\bar{\boldsymbol{\theta}})]^{-1} \nabla R_n(\boldsymbol{\theta}^*)$$

for some $\bar{\boldsymbol{\theta}}$ between $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}_n$. On the event where $\inf_{\boldsymbol{\theta}\in B_{\delta_0}(\boldsymbol{\theta}^*)} \lambda_{\min}(\nabla^2 R_n(\boldsymbol{\theta})) \geq \lambda_0/2$ and $\|\nabla R_n(\boldsymbol{\theta}^*)\|_2 < \frac{\lambda_0 t}{2}\sqrt{d/n}$, we have

$$\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|_2 \leq \frac{2}{\lambda_0}\|\nabla R_n(\boldsymbol{\theta}^*)\|_2 < t\sqrt{d/n}.$$

To ensure $\hat{\boldsymbol{\theta}}_n \in B_{\delta_0}(\boldsymbol{\theta}^*)$, we let $t\sqrt{d/n} \leq \delta_0$. Combining with previous bounds yield

$$\mathbb{P}\left(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|_2 \geq t\sqrt{d/n}\right) \leq 2d\exp\left(-\frac{\lambda_0^2 t^2}{8\sigma^2 L^2}\right) + 2d\exp\left(-\frac{\lambda_0^2 n}{512L^4}\right) + 2d\exp\left(-\frac{\lambda_0^2 n}{512\sigma^2 B^2}\right).$$

Finally, we relate the function estimation error to the parameter estimation error:

$$\|\hat{v} - v^*\|_\infty = \sup_{x\in\mathcal{X}} |g(\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}) - g(\boldsymbol{\theta}^*, \boldsymbol{x})| \leq \sup_{x\in\mathcal{X}} \|\nabla g(\tilde{\boldsymbol{\theta}}, \boldsymbol{x})\|_2 \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|_2 \leq L\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|_2$$

for some $\tilde{\boldsymbol{\theta}}$ between $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}_n$. Therefore, we have

$$\begin{aligned}
&\mathbb{P}\left(\|\hat{v} - v^*\|_\infty \geq t\sqrt{d/n}\right) \leq \mathbb{P}\left(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|_2 \geq \frac{t}{L}\sqrt{d/n}\right) \\
&\leq 2d\exp\left(-\frac{\lambda_0^2 t^2}{8\sigma^2 L^4}\right) + 2d\exp\left(-\frac{\lambda_0^2 n}{512L^4}\right) + 2d\exp\left(-\frac{\lambda_0^2 n}{512\sigma^2 B^2}\right).
\end{aligned}$$

Set $t = \sqrt{\frac{1}{c_1}\ln(8d/\delta)}$ and $n \geq \frac{1}{c_2}\ln(8d/\delta)$, we have:

$$\begin{aligned}
&2d\exp(-c_1 t^2) = 2d \cdot \frac{\delta}{8d} = \delta/4, \\
&2d\exp\left(-\frac{\lambda_0^2 n}{512L^4}\right) \leq 2d\exp(-c_2 n) \leq 2d \cdot \frac{\delta}{8d} = \delta/4, \\
&2d\exp\left(-\frac{\lambda_0^2 n}{512\sigma^2 B^2}\right) \leq 2d\exp(-c_2 n) \leq \delta/4.
\end{aligned}$$

Summing over all probabilities, we obtain:

$$\mathbb{P}\left(\|\hat{v} - v^*\|_\infty \geq \sqrt{\frac{d}{c_1 n}\ln\left(\frac{8d}{\delta}\right)}\right) \leq \delta/4 + \delta/4 + \delta/4 = 3\delta/4 \leq \delta,$$

which completes the proof. □

From the previous theorem, we know that Assumption 4 holds for the parametric function spaces with $\rho(\delta) = \mathcal{O}(d\ln(d/\delta))$ and offline least squares regression oracle. A special case is the linear function space, where $g(\boldsymbol{\theta}, \boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x}$, which is widely used in the literature.

A.2 Finite Function Space

We now discuss a second example of an offline regression oracle in the case of finite function spaces.

Theorem 4. *Consider a finite function space $\mathcal{V}$ with $|\mathcal{V}| < \infty$, and a target function $v^* \in \mathcal{V}$. Assume there exists $B > 0$ such that $\sup_{v\in\mathcal{V},\boldsymbol{x}\in\mathcal{X}} |v(\boldsymbol{x})| \le B$. Given a dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ of size n generated as $y_i = v^*(\boldsymbol{x}_i) + \epsilon_i$, where:*

1. *$\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. random variables on $\mathcal{X}$,*
2. *$\{\epsilon_i\}_{i=1}^n$ are i.i.d. random noise with $\mathbb{E}[\epsilon_i] = 0$ and $|\epsilon_i| \le B$ a.s.,*
3. *All $\boldsymbol{x}_i$ and ϵ_i are mutually independent,*
4. *The minimum expected gap satisfies $\mu_{\min} = \min_{v\neq v^*} \mathbb{E}_x[(v^*(\boldsymbol{x}) - v(\boldsymbol{x}))^2] > 0$.*

Then the estimator $\hat{v} = \arg\min_{v\in\mathcal{V}} \sum_{i=1}^n (y_i - v(\boldsymbol{x}_i))^2$ satisfies

$$\mathbb{P}(v^* \neq \hat{v}) \le (|\mathcal{V}| - 1) \exp\left(-\frac{n\mu_{\min}^2}{128B^4}\right). \tag{8}$$

Proof. The least squares regression oracle uses the following estimator:

$$\hat{v} = \operatorname*{argmin}_{v\in\mathcal{V}} \sum_{i=1}^n (y_i - v(\boldsymbol{x}_i))^2.$$

Define $A = \{\hat{v} \neq v^*\}$. The boundedness implies $\sup_x |\hat{v}(x) - v^*(\boldsymbol{x})| \le 2B$. Then the oracle selects $v \neq v^*$ if

$$\sum_{i=1}^n (y_i - v(\boldsymbol{x}_i))^2 \le \sum_{i=1}^n (y_i - v^*(\boldsymbol{x}_i))^2.$$

Substituting $y_i = v^*(\boldsymbol{x}_i) + \epsilon_i$ and simplifying:

$$\sum_{i=1}^n \left[(v^*(\boldsymbol{x}_i) - v(\boldsymbol{x}_i))^2 + 2(v^*(\boldsymbol{x}_i) - v(\boldsymbol{x}_i))\epsilon_i\right] \le 0.$$

Let $d_v(\boldsymbol{x}_i) = v^*(\boldsymbol{x}_i) - v(\boldsymbol{x}_i)$. Then:

$$S_v = \sum_{i=1}^n \left[d_v(\boldsymbol{x}_i)^2 + 2d_v(\boldsymbol{x}_i)\epsilon_i\right] \le 0.$$

By the union bound:

$$\mathbb{P}(A) \le \sum_{v\neq v^*} \mathbb{P}(S_v \le 0).$$

For each $v \neq v^*$, we have $\mathbb{E}[S_v] = \mathbb{E}\left[\sum_{i=1}^n d_v(\boldsymbol{x}_i)^2\right] = n\mu_v \ge n\mu_{\min} > 0$. Since $|d_v(\boldsymbol{x}_i)| \le 2B$ and $|\epsilon_i| \le B$, we have

$$|d_v(\boldsymbol{x}_i)^2 + 2d_v(\boldsymbol{x}_i)\epsilon_i| \le (2B)^2 + 2(2B)(B) = 8B^2.$$

The Hoeffding's inequality yields

$$\mathbb{P}(S_v \le 0) \le \exp\left(-\frac{2(\mathbb{E}[S_v])^2}{n(16B^2)^2}\right) = \exp\left(-\frac{(\mathbb{E}[S_v])^2}{128nB^4}\right) \le \exp\left(-\frac{n\mu_{\min}^2}{128B^4}\right).$$

Therefore, we have:

$$\mathbb{P}(A) \le (|\mathcal{V}| - 1)\exp\left(-\frac{n\mu_{\min}^2}{128B^4}\right).$$

□

To bound the right-hand-side of (8) by δ, we set the sample size n

$$n \ge \frac{128B^4}{\mu_{\min}^2}\ln(|\mathcal{V}|/\delta).$$

so that

$$(|\mathcal{V}| - 1)\exp\left(-\frac{n\mu_{\min}^2}{128B^4}\right) \le \delta.$$

Though Theorem 4 does not fully satisfy Assumption 4, it can still works for Theorem 1 as it provides the sample complexity of learning in finite function spaces.

B Proofs

B.1 Proofs for Section 3

Lemma 8 (Azuma's Inequality)**.** *Let $\{X_\tau\}_{\tau=1}^n$ be a martingale difference sequence with respect to a filtration $\{\mathcal{F}_\tau\}_{\tau=0}^n$, i.e., $\mathbb{E}[X_\tau \mid \mathcal{F}_{\tau-1}] = 0$ for all τ. Assume $a_\tau \le X_\tau \le b_\tau$ a.s. for $\tau = 1, \dots, n$. Then for any $\iota \in (0,1)$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{\tau=1}^n X_\tau\right| \ge \frac{1}{n}\sqrt{\frac{1}{2}\ln(2/\iota)\sum_{\tau=1}^n (b_\tau - a_\tau)^2}\right) \le \iota.$$

Lemma 1. *Fix an episode k. For any round t in the episode k, and any layer s at round t, with probability at least $1 - \delta/(S_k N_k T_k)$, each $j \in [N_k]$ satisfies*

$$|\xi_j^* - w_{t,s}^j| \le r_{t,s}^j + L\eta_{t,s}^j, \quad \textit{where} \quad \eta_{t,s}^j = \frac{1}{|\Psi_t^s(j)|}\sum_{\tau \in \Psi_t^s(j)} |\hat{v}_k(\boldsymbol{x}_\tau) - v^*(\boldsymbol{x}_\tau)|.$$

Proof. Let $\mathcal{H}_0$ denote the filtration generated by

$$\left\{(\tau, \boldsymbol{x}_\tau, p_\tau), \tau \in \bigcup_{s' \le s} \Psi_t^{s'}\right\} \cup \left\{y_\tau, \tau \in \bigcup_{s' < s} \Psi_t^{s'}\right\}.$$

By construction of the layered partition, the event $\{\tau \in \Psi_t^s\}$ is $\mathcal{H}_0$-measurable and does not depend on $\{y_\tau : \tau \in \Psi_t^s\}$; since $\{\epsilon_\tau\}_\tau$ are i.i.d. and independent of contexts, conditioning on $\mathcal{H}_0$ renders

$\{\epsilon_\tau : \tau \in \Psi_t^s\}$ independent. Considering the conditional independence of ϵ_τ indexed in Ψ_t^s, we can deduce that for $\tau \in \Psi_t^s(j)$

$$\begin{aligned}
&\mathbb{E}[y_\tau \mid \{y_{\tau'} : \tau' \in \Psi_t^s(j), \tau' < \tau\}, \mathcal{H}_0] \\
&= \mathbb{E}[\mathbb{I}\{\epsilon_\tau > m_j + \hat{v}_k(\boldsymbol{x}_\tau) - v^*(\boldsymbol{x}_\tau)\} \mid \{\epsilon_{\tau'} : \tau' \in \Psi_t^s(j), \tau' < \tau\}, \mathcal{H}_0] \\
&= \mathbb{E}[\mathbb{I}\{\epsilon_\tau > m_j + \hat{v}_k(\boldsymbol{x}_\tau) - v^*(\boldsymbol{x}_\tau)\} \mid \mathcal{H}_0] \\
&= 1 - F(m_j + \hat{v}_k(\boldsymbol{x}_\tau) - v^*(\boldsymbol{x}_\tau)).
\end{aligned}$$

Now we consider the concentration for $y_\tau, \tau \in \Psi_t^s(j)$. For notation brevity, denote $\mathcal{H}_\tau$ as the filtration generated by $\{y_{\tau'} : \tau' \in \Psi_t^s(j), \tau' < \tau\} \cup \mathcal{H}_0$. Then for $\tau = 0$, the definition of $\mathcal{H}_\tau$ is consistent with $\mathcal{H}_0$. To apply Azuma's inequality (Lemma 8) with $b_\tau = 1$ and $a_\tau = 0$, it is easy to check $y_\tau - \mathbb{E}[y_\tau|\mathcal{H}_\tau]$ is a martingale difference sequence adapted to filtration $\mathcal{H}_\tau$. Hence, we can derive the following results:

$$\mathbb{P}\left(\left|w_{t,s}^j - \frac{\sum_{\tau\in\Psi_t^s(j)} \mathbb{E}[y_\tau|\mathcal{H}_\tau]}{|\Psi_t^s(j)|}\right| \geq \sqrt{\frac{2\ln(2S_kN_kT_k/\delta)}{|\Psi_t^s(j)|}}\right) \leq \frac{\delta}{S_kN_kT_k}.$$

Hence, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\left|w_{t,s}^j - \frac{\sum_{\tau\in\Psi_t^s(j)} \mathbb{E}[y_\tau|\mathcal{H}_\tau]}{|\Psi_t^s(j)|}\right| \geq r_{t,s}^j\right) \\
&\leq \mathbb{P}\left(\left|w_{t,s}^j - \frac{\sum_{\tau\in\Psi_t^s(j)} \mathbb{E}[y_\tau|\mathcal{H}_\tau]}{|\Psi_t^s(j)|}\right| \geq \sqrt{\frac{2\ln(2S_kN_kT_k/\delta)}{|\Psi_t^s(j)|}}\right) \\
&\leq \frac{\delta}{S_kN_kT_k}.
\end{aligned}$$

From Assumption 3, we know

$$|\xi_j^* - \mathbb{E}[y_\tau \mid \mathcal{H}_\tau]| = |(1 - F(m_j)) - (1 - F(m_j + \hat{v}_k(\boldsymbol{x}_\tau) - v^*(\boldsymbol{x}_\tau)))| \leq L|\hat{v}_k(\boldsymbol{x}_\tau) - v^*(\boldsymbol{x}_\tau)|.$$

Therefore, we have

$$\left|\frac{\sum_{\tau\in\Psi_t^s(j)}(\xi_j^* - \mathbb{E}[y_\tau|\mathcal{H}_\tau])}{|\Psi_t^s(j)|}\right| \leq \frac{L\sum_{\tau\in\Psi_t^s(j)} |\hat{v}_k(\boldsymbol{x}_\tau) - v^*(\boldsymbol{x}_\tau)|}{|\Psi_t^s(j)|},$$

so by the triangle inequality, we find

$$\begin{aligned}
&\mathbb{P}\left(\left|w_{t,s}^j - \xi_j^*\right| \geq r_{t,s}^j + \frac{L\sum_{\tau\in\Psi_t^s(j)} |\hat{v}_k(\boldsymbol{x}_\tau) - v^*(\boldsymbol{x}_\tau)|}{|\Psi_t^s(j)|}\right) \\
\leq &\mathbb{P}\left(\left|w_{t,s}^j - \frac{\sum_{\tau\in\Psi_t^s(j)} \mathbb{E}[y_\tau|\mathcal{H}_\tau]}{|\Psi_t^s(j)|}\right| + \left|\frac{\sum_{\tau\in\Psi_t^s(j)}(\xi_j^* - \mathbb{E}[y_\tau|\mathcal{H}_\tau])}{|\Psi_t^s(j)|}\right|\right. \\
&\qquad\qquad \left. \geq r_{t,s}^j + \frac{L\sum_{\tau\in\Psi_t^s(j)} |\hat{v}_k(\boldsymbol{x}_\tau) - v^*(\boldsymbol{x}_\tau)|}{|\Psi_t^s(j)|}\right) \\
\leq &\mathbb{P}\left(\left|w_{t,s}^j - \frac{\sum_{\tau\in\Psi_t^s(j)} \mathbb{E}[y_\tau|\mathcal{H}_\tau]}{|\Psi_t^s(j)|}\right| \geq r_{t,s}^j\right) \\
\leq &\frac{\delta}{S_kN_kT_k}.
\end{aligned}$$

□

B.2 Proofs for Upper Bounds in Section 4.1

Recall that

$$\tilde{p}_t^* \triangleq m_{j_t^*} + \hat{v}_k(\boldsymbol{x}_t), \quad \text{where} \quad j_t^* = \operatorname*{argmax}_{j \in [N_k]} \mathsf{Rev}_t\left(m_j + \hat{v}_k(\boldsymbol{x}_t)\right),$$
$$\tilde{p}_{t,s}^* \triangleq m_{j_{t,s}^*} + \hat{v}_k(\boldsymbol{x}_t), \quad \text{where} \quad j_{t,s}^* = \operatorname*{argmax}_{j \in \mathcal{A}_{t,s}} \mathsf{Rev}_t(m_j + \hat{v}_k(\boldsymbol{x}_t)).$$

Lemma 2. *For each round t in episode k and each layer $s \in [s_t]$, conditional on event Γ_k, we have*

$$\mathsf{Rev}_t(\tilde{p}_t^*) - \mathsf{Rev}_t(\tilde{p}_{t,s}^*) \le 4BL(s-1)\|\hat{v}_k - v^*\|_\infty.$$

Proof. We prove this lemma by induction on s. For $s = 1$, the lemma holds naturally since $\mathcal{A}_{t,1} = [N_k]$ and hence $j_t^* = j_{t,1}^*$. Assume that the bound holds at the layer $s \le s_t - 1$. It suffices to show that

$$\mathsf{Rev}_t\left(m_{j_{t,s}^*} + \hat{v}_k(\boldsymbol{x}_t)\right) - \mathsf{Rev}_t\left(m_{j_{t,s+1}^*} + \hat{v}_k(\boldsymbol{x}_t)\right) \le 4BL\|\hat{v}_k - v^*\|_\infty.$$

If $j_{t,s}^* = j_{t,s+1}^*$, then the desired bound holds. Hence we assume that $j_{t,s}^* \notin \mathcal{A}_{t,s+1}$. Let

$$\hat{j}_{t,s} := \operatorname*{argmax}_{j \in \mathcal{A}_{t,s}} \left(m_j + \hat{v}_k(\boldsymbol{x}_t)\right)\left(w_{t,s}^j + r_{t,s}^j\right)$$

be the index with the highest UCB in $\mathcal{A}_{t,s}$. From Step 14 of Algorithm 2, we know that $\hat{j}_{t,s} \in \mathcal{A}_{t,s+1}$. Then we have

$$\begin{aligned}
&\mathsf{Rev}_t\left(m_{j_{t,s}^*} + \hat{v}_k(\boldsymbol{x}_t)\right) - \mathsf{Rev}_t\left(m_{j_{t,s+1}^*} + \hat{v}_k(\boldsymbol{x}_t)\right) \\
&\le \mathsf{Rev}_t\left(m_{j_{t,s}^*} + \hat{v}_k(\boldsymbol{x}_t)\right) - \mathsf{Rev}_t\left(m_{\hat{j}_{t,s}} + \hat{v}_k(\boldsymbol{x}_t)\right) \\
&\le \left(m_{j_{t,s}^*} + \hat{v}_k(\boldsymbol{x}_t)\right)\left(1 - F\left(m_{j_{t,s}^*}\right)\right) - \left(m_{\hat{j}_{t,s}} + \hat{v}_k(\boldsymbol{x}_t)\right)\left(1 - F\left(m_{\hat{j}_{t,s}}\right)\right) \\
&\quad + 2BL\left|\hat{v}_k(\boldsymbol{x}_t) - v^*(\boldsymbol{x}_t)\right|.
\end{aligned}$$

From the definition of Γ_k in (3), we know that

$$\begin{aligned}
&\left(m_{j_{t,s}^*} + \hat{v}_k(\boldsymbol{x}_t)\right)\left(1 - F\left(m_{j_{t,s}^*}\right)\right) - \left(m_{\hat{j}_{t,s}} + \hat{v}_k(\boldsymbol{x}_t)\right)\left(1 - F\left(m_{\hat{j}_{t,s}}\right)\right) \\
\le\ &\left(m_{j_{t,s}^*} + \hat{v}_k(\boldsymbol{x}_t)\right)\left(w_{t,s}^{j_{t,s}^*} + r_{t,s}^{j_{t,s}^*}\right) - \left(m_{\hat{j}_{t,s}} + \hat{v}_k(\boldsymbol{x}_t)\right)\left(w_{t,s}^{\hat{j}_{t,s}} - r_{t,s}^{\hat{j}_{t,s}}\right) + BL\eta_{t,s}^{j_{t,s}^*} + BL\eta_{t,s}^{\hat{j}_{t,s}} \\
=\ &\left(m_{j_{t,s}^*} + \hat{v}_k(\boldsymbol{x}_t)\right)\left(w_{t,s}^{j_{t,s}^*} + r_{t,s}^{j_{t,s}^*}\right) - \left(m_{\hat{j}_{t,s}} + \hat{v}_k(\boldsymbol{x}_t)\right)\left(w_{t,s}^{\hat{j}_{t,s}} + r_{t,s}^{\hat{j}_{t,s}}\right) + 2\left(m_{\hat{j}_{t,s}} + \hat{v}_k(\boldsymbol{x}_t)\right) r_{t,s}^{\hat{j}_{t,s}} \\
&\quad + BL\eta_{t,s}^{j_{t,s}^*} + BL\eta_{t,s}^{\hat{j}_{t,s}}.
\end{aligned}$$

From the statistical precision check (Step 13 of Algorithm 2), we know

$$\left(m_{\hat{j}_{t,s}} + \hat{v}_k(\boldsymbol{x}_t)\right) r_{t,s}^{\hat{j}_{t,s}} \le B2^{-s}$$

as $s < s_t$. Furthermore, since $j_{t,s}^* \notin \mathcal{A}_{t,s+1}$, the elimination step also implies that

$$\begin{aligned}
\left(m_{j_{t,s}^*} + \hat{v}_k(\boldsymbol{x}_t)\right)\left(w_{t,s}^{j_{t,s}^*} + r_{t,s}^{j_{t,s}^*}\right) - \left(m_{\hat{j}_{t,s}} + \hat{v}_k(\boldsymbol{x}_t)\right)\left(w_{t,s}^{\hat{j}_{t,s}} + r_{t,s}^{\hat{j}_{t,s}}\right) &< -B2^{1-s} \\
&\le -2\left(m_{\hat{j}_{t,s}} + \hat{v}_k(\boldsymbol{x}_t)\right) r_{t,s}^{\hat{j}_{t,s}}.
\end{aligned}$$

Combining all the inequalities above, we obtain

$$\mathsf{Rev}_t\left(m_{j^*_{t,s}}+\hat{v}_k(\boldsymbol{x}_t)\right)-\mathsf{Rev}_t\left(m_{j^*_{t,s+1}}+\hat{v}_k(\boldsymbol{x}_t)\right)\le 2BL\left|\hat{v}_k(\boldsymbol{x}_t)-v^*(\boldsymbol{x}_t)\right|+BL\eta_{t,s}^{j^*_{t,s}}+BL\eta_{t,s}^{\hat{j}_{t,s}}.$$

Recall that $\eta_{t,s}^{j}\le\|\hat{v}_k-v^*\|_\infty$ for any t,s,j, combining all the above inequalities yields the desired inequality. □

Lemma 3. *For each round t in episode k and each layer $2\le s\le s_t$, conditional on event Γ_k, we have*

$$\mathsf{Rev}_t(\tilde{p}^*_{t,s})-\mathsf{Rev}_t(p_t)\le 8B\cdot 2^{-s}+4BL\|\hat{v}_k-v^*\|_\infty.$$

Proof. For all $2\le s\le s_t$, Step 14 of Algorithm 2 shows that

$$(m_j+\hat{v}_k(\boldsymbol{x}_t))(w_{t,s-1}^{j}+r_{t,s-1}^{j})\ge(m_{j^*_{t,s}}+\hat{v}_k(\boldsymbol{x}_t))(w_{t,s-1}^{j^*_{t,s}}+r_{t,s-1}^{j^*_{t,s}})-B2^{2-s},\forall j\in\mathcal{A}_{t,s},$$

as $j^*_{t,s}\in\mathcal{A}_{t,s}\subset\mathcal{A}_{t,s-1}$. Furthermore, Step 13 of Algorithm 2 implies that $(m_j+\hat{v}_k(\boldsymbol{x}_t))r_{t,s-1}^{j}\le B2^{1-s}$ for all $j\in\mathcal{A}_{t,s-1}$. Combining two inequalities, we obtain

$$\begin{aligned}
&(m_{j^*_{t,s}}+\hat{v}_k(\boldsymbol{x}_t))(w_{t,s-1}^{j^*_{t,s}}+r_{t,s-1}^{j^*_{t,s}})-(m_j+\hat{v}_k(\boldsymbol{x}_t))(w_{t,s-1}^{j}-r_{t,s-1}^{j})\\
\le&2(m_j+\hat{v}_k(\boldsymbol{x}_t))r_{t,s-1}^{j}+B2^{2-s}\\
\le&4B2^{1-s},\quad\forall j\in\mathcal{A}_{t,s}.
\end{aligned}$$

Therefore, from the definition of Γ_k in (3), we have

$$\begin{aligned}
4B2^{1-s}&\ge(m_{j^*_{t,s}}+\hat{v}_k(\boldsymbol{x}_t))(w_{t,s-1}^{j^*_{t,s}}+r_{t,s-1}^{j^*_{t,s}})-(m_{j_t}+\hat{v}_k(\boldsymbol{x}_t))(w_{t,s-1}^{j_t}-r_{t,s-1}^{j_t})\\
&\ge\mathsf{Rev}_t\left(\tilde{p}^*_{t,s}\right)-\mathsf{Rev}_t\left(p_t\right)-BL\eta_{t,s-1}^{j^*_{t,s}}-BL\eta_{t,s-1}^{j_t}-2BL\|\hat{v}_k-v^*\|_\infty.
\end{aligned}$$

Recall that $\eta_{t,s}^{j}\le\|\hat{v}_k-v^*\|_\infty$ for any t,s,j, combining all the above inequalities yields the desired inequality. □

Lemma 4. *For each round t in episode k and each layer $2\le s\le s_t$, conditional on event Γ_k, we have*

$$\mathsf{Rev}_t(\tilde{p}^*_{t})-\mathsf{Rev}_t(p_t)\le 8B\cdot 2^{-s}+4BLs\|\hat{v}_k-v^*\|_\infty.$$

Proof. It follows from Lemma 2 and Lemma 3 that

$$\begin{aligned}
&\mathsf{Rev}_t(m_{j^*_t}+\hat{v}_k(\boldsymbol{x}_t))-\mathsf{Rev}_t(m_{j_t}+\hat{v}_k(\boldsymbol{x}_t))\\
=&\mathsf{Rev}_t(m_{j^*_{t,s}}+\hat{v}_k(\boldsymbol{x}_t))-\mathsf{Rev}_t(m_{j_t}+\hat{v}_k(\boldsymbol{x}_t))+\mathsf{Rev}_t(m_{j^*_t}+\hat{v}_k(\boldsymbol{x}_t))-\mathsf{Rev}_t(m_{j^*_{t,s}}+\hat{v}_k(\boldsymbol{x}_t))\\
\le&8B\cdot 2^{-s}+4BLs\|\hat{v}_k-v^*\|_\infty.
\end{aligned}$$

□

Recall that $\Psi_{T_k+1}^{S_k}$ is the set of rounds chosen during the exploitation step (Step 8 of Algorithm 2), i.e., time steps with $s_t=S_k$.

Lemma 9. *Conditional on event Γ_k, for every round $t \in \Psi^{S_k}_{T_k+1}$ in the episode k, we have*

$$\mathsf{Rev}_t(\tilde{p}^*_t) - \mathsf{Rev}_t(p_t) \le \frac{4B}{\sqrt{T_k}} + 6BL\|\hat{v}_k - v^*\|_\infty \ln T_k.$$

Proof. Since the stopping layer $s_t = S_k$, it follows from Assumption 3 and Lemma 2 that

$$\begin{aligned}
&\mathsf{Rev}_t(\tilde{p}^*_t) - \mathsf{Rev}_t(p_t) \\
&= \mathsf{Rev}_t(\tilde{p}^*_t) - \mathsf{Rev}_t(\tilde{p}^*_{t,S_k}) + \mathsf{Rev}_t(\tilde{p}^*_{t,S_k}) - \mathsf{Rev}_t(p_t) \\
&\le 4BL\|\hat{v}_k - v^*\|_\infty (S_k - 1) + \mathsf{Rev}_t(m_{j^*_{t,S_k}} + \hat{v}_k(\boldsymbol{x}_t)) - \mathsf{Rev}_t(m_{j_t} + \hat{v}_k(\boldsymbol{x}_t)) \\
&\le (m_{j^*_{t,S_k}} + \hat{v}_k(\boldsymbol{x}_t))(w^{j^*_t}_{t,S_k} + r^{j^*_t}_{t,S_k}) - (m_{j_t} + \hat{v}_k(\boldsymbol{x}_t))(w^{j_t}_{t,S_k} - r^{j_t}_{t,S_k}) + 6BL\|\hat{v}_k - v^*\|_\infty \ln T_k \\
&\le (m_{j^*_{t,S_k}} + \hat{v}_k(\boldsymbol{x}_t))(w^{j^*_t}_{t,S_k} + r^{j^*_t}_{t,S_k}) - (m_{j_t} + \hat{v}_k(\boldsymbol{x}_t))(w^{j_t}_{t,S_k} + r^{j_t}_{t,S_k}) \\
&\qquad + \frac{4B}{\sqrt{T_k}} + 6BL\|\hat{v}_k - v^*\|_\infty \ln T_k \\
&\le \frac{4B}{\sqrt{T_k}} + 6BL\|\hat{v}_k - v^*\|_\infty \ln T_k.
\end{aligned}$$

The last inequality comes from the fact that j_t in $\mathcal{A}_{t,S_k}$ is the index corresponding to the largest UCB.

□

Lemma 10. *Assuming $|\Psi^s_{T_k+1}| \ge 1$, we have*

$$\sum_{t \in \Psi^s_{T_k+1}} p_t r^{j_t}_{t,s} \le 2B\sqrt{2N_k |\Psi^s_{T_k+1}| \ln(2S_k N_k T_k/\delta)}.$$

Proof. We have

$$\sum_{t \in \Psi^s_{T_k+1}} p_t r^{j_t}_{t,s} = \sqrt{2\ln(2S_k N_k T_k/\delta)} \sum_{t \in \Psi^s_{T_k+1}} \frac{p_t}{\sqrt{|\Psi^s_t(j_t)|}}.$$

It suffices to bound the term $\sum_{t \in \Psi^s_{T_k+1}} \frac{p_t}{\sqrt{|\Psi^s_t(j_t)|}}$. Notice that

$$\begin{aligned}
\sum_{t \in \Psi^s_{T_k+1}} \frac{p_t}{\sqrt{|\Psi^s_t(j_t)|}} &\le \sum_{j=1}^{N_k} \sum_{t=1}^{|\Psi^s_{T_k+1}(j)|} \frac{B}{\sqrt{t}} \\
&\le \sum_{j=1}^{N_k} 2B\sqrt{|\Psi^s_{T_k+1}(j)|} \\
&\le 2B\sqrt{N_k \sum_{j=1}^{N_k} |\Psi^s_{T_k+1}(j)|} \\
&\le 2B\sqrt{N_k |\Psi^s_{T_k+1}|}.
\end{aligned}$$

Therefore, we conclude that

$$\sum_{t \in \Psi^s_{T_k+1}} p_t r^{j_t}_{t,s} \le 2B\sqrt{2N_k |\Psi^s_{T_k+1}| \ln(2S_k N_k T_k/\delta)}.$$

□

Lemma 11. *For all $s < S_k$, we have*

$$|\Psi^s_{T_k+1}| \le 2^{s+1}\sqrt{2N_k|\Psi^s_{T_k+1}|\ln(2S_kN_kT_k/\delta)}.$$

Proof. For any $s < S_k$, the data enters the stopping layer s only during the exploration step (Step 12 of Algorithm 2). In this case, we know that

$$\sum_{t\in\Psi^s_{T_k+1}} p_t r^{j_t}_{t,s} \ge B2^{-s}|\Psi^s_{T_k+1}|.$$

By Lemma 10, we obtain

$$\sum_{t\in\Psi^s_{T_k+1}} p_t r^{j_t}_{t,s} \le 2B\sqrt{2N_k|\Psi^s_{T_k+1}|\ln(2S_kN_kT_k/\delta)}.$$

Therefore, combing above inequalities, we have

$$|\Psi^s_{T_k+1}| \le 2^{s+1}\sqrt{2N_k|\Psi^s_{T_k+1}|\ln(2S_kN_kT_k/\delta)}.$$

□

Lemma 5. *Under Assumptions 1, 2 and 3, the learning regret $\sum_{t=1}^T \mathcal{R}^1_t$ is bounded by*

$$\sum_{k=1}^{\lceil\log_2 T\rceil}\Big[16B\sqrt{2N_kT_k\ln(2S_kT_kN_k/\delta)\ln T_k} + 9BL\|\hat{v}_k - v^*\|_\infty T_k\ln T_k + 4BT_k^{\frac{1}{2}} + 64BN_k\ln(2S_kT_kN_k/\delta)\Big] \quad \textit{with probability at least } 1-\lceil\log_2 T\rceil\delta.$$

Proof. We first consider a fixed episode k and assume that event Γ_k holds.

1. **Rounds in the terminal layer S_k.** By definition, $\Psi^{S_k}_{T_k+1}$ is the set of rounds whose stopping layer is S_k; in this layer we have $(m_j + \hat{v}_k(\boldsymbol{x}_t))r^j_{t,S_k} \le 2B/\sqrt{T_k}$. By Lemma 9,

$$\sum_{t\in\Psi^{S_k}_{T_k+1}} \mathcal{R}^1_t \le \Big(\frac{4B}{\sqrt{T_k}} + 6BL\|\hat{v}_k - v^*\|_\infty \ln T_k\Big)T_k \le 4B\sqrt{T_k} + 6BL\|\hat{v}_k - v^*\|_\infty T_k\ln T_k.$$

2. **Rounds in layers $s = 2,\dots,S_k-1$.** Recall that $\tilde{p}^*_t = \arg\max_{j\in[N_k]} \mathsf{Rev}_t\big(m_j + \hat{v}_k(\boldsymbol{x}_t)\big)$ is the discrete empirical best price. Summing the per-round bound from Lemma 4 over $t \in \Psi^s_{T_k+1}$ and then over $s = 2,\dots,S_k-1$ gives

$$\sum_{s=2}^{S_k-1}\sum_{t\in\Psi^s_{T_k+1}} [\mathsf{Rev}_t(\tilde{p}^*_t) - \mathsf{Rev}_t(p_t)] \le \sum_{s=2}^{S_k-1}\Big(8B\cdot 2^{-s} + (4s-2)BL\|\hat{v}_k - v^*\|_\infty\Big)|\Psi^s_{T_k+1}|.$$

For the first term, apply Lemma 11 and Cauchy-Schwarz:

$$\sum_{s=2}^{S_k-1} 8B\cdot 2^{-s}|\Psi^s_{T_k+1}| \le \sum_{s=2}^{S_k-1} 16B\sqrt{2N_k|\Psi^s_{T_k+1}|\ln(2S_kN_kT_k/\delta)}$$
$$\le 16B\sqrt{2N_k\ln(2S_kN_kT_k/\delta)}\sqrt{(S_k-2)\sum_{s=2}^{S_k-1}|\Psi^s_{T_k+1}|}.$$

Since $\sum_{s=1}^{S_k} |\Psi^s_{T_k+1}| = T_k$ and $S_k = \lceil \frac{1}{2}\log_2 T_k \rceil \le \frac{1}{2}\log_2 T_k + 1$, we have $S_k - 2 \le \frac{1}{2}\log_2 T_k \le (\ln T_k)/(2\ln 2) \le \ln T_k$. Therefore,

$$\sum_{s=2}^{S_k-1} 8B \cdot 2^{-s} |\Psi^s_{T_k+1}| \le 16B\sqrt{2N_kT_k \ln(2S_kN_kT_k/\delta)\ln T_k}.$$

For the second term, use the crude bound $\sum_{s=2}^{S_k-1}(4s-2)|\Psi^s_{T_k+1}| \le (4S_k-6)\sum_{s=2}^{S_k-1}|\Psi^s_{T_k+1}| \le (4S_k-6)T_k$, and since $S_k \le \frac{1}{2}\log_2 T_k + 1$,

$$4S_k - 6 \le 2\log_2 T_k - 2 = \tfrac{2}{\ln 2}\ln T_k - 2 \le 3\ln T_k - 2.$$

Hence

$$\sum_{s=2}^{S_k-1}(4s-2)BL\|\hat{v}_k - v^*\|_\infty |\Psi^s_{T_k+1}| \le BL\|\hat{v}_k - v^*\|_\infty T_k(3\ln T_k - 2).$$

3. **Rounds in the first layer $s = 1$.** By Lemma 11 with $s = 1$, $|\Psi^1_{T_k+1}| \le 32N_k \ln(2S_kN_kT_k/\delta)$. Using the trivial per-round bound $\mathcal{R}^1_t \le 2B$ on this layer,

$$\sum_{t\in\Psi^1_{T_k+1}} \mathcal{R}^1_t \le 64BN_k \ln\big(2S_kN_kT_k/\delta\big).$$

Since $\bigcup_{s=1}^{S_k}\Psi^s_{T_k+1} = [T_k]$, adding the contributions from all rounds yields, on Γ_k,

$$\begin{aligned}\sum_{t=2^{k-1}+T^e_k+1}^{2^k} \mathcal{R}^1_t \le{}& 16B\sqrt{2N_kT_k\ln(2S_kN_kT_k/\delta)\ln T_k} + BL\|\hat{v}_k - v^*\|_\infty T_k(3\ln T_k - 2)\\ &+ 4B\sqrt{T_k} + 6BL\|\hat{v}_k - v^*\|_\infty T_k \ln T_k + 64BN_k\ln\big(2S_kN_kT_k/\delta\big).\end{aligned}$$

Absorbing the harmless "-2" into the logarithmic factor and summing over $k = 1, \dots, \lceil\log_2 T\rceil$, then applying a union bound over $\{\Gamma_k\}$ gives the stated episode-wise sum and the overall probability at least $1 - \lceil\log_2 T\rceil\delta$. □

Lemma 6. *Under Assumptions 1 and 2, the discretization regret $\sum_{t=1}^T \mathcal{R}^2_t$ is upper bounded by*

$$\sum_{t=1}^T \mathcal{R}^2_t \le \sum_{k=1}^{\lceil\log_2 T\rceil} \frac{3BT_k}{N_k}.$$

Proof. Recall that p^*_t is the (continuous) optimal price, $\tilde{p}^*_t$ is the discrete best price among the candidate set, and m_j are the *midpoints* of an equi-spaced grid partitioning $[-\|\hat{v}_k\|_\infty, B + \|\hat{v}_k\|_\infty]$ into N_k subintervals. Define the left neighbor of p^*_t on the discrete grid by

$$\dot{p}_t \triangleq \max\Big\{0, \max_{j: m_j + \hat{v}_k(\boldsymbol{x}_t) \le p^*_t} \big(m_j + \hat{v}_k(\boldsymbol{x}_t)\big)\Big\}.$$

Since $\tilde{p}^*_t$ maximizes the discrete revenue, we have

$$\tilde{p}^*_t\big(1 - F(\tilde{p}^*_t - v^*(\boldsymbol{x}_t))\big) \ge \dot{p}_t\big(1 - F(\dot{p}_t - v^*(\boldsymbol{x}_t))\big),$$

and therefore

$$\begin{aligned}
&p_t^*(1-F(p_t^*-v^*(\boldsymbol{x}_t)))-\tilde{p}_t^*(1-F(\tilde{p}_t^*-v^*(\boldsymbol{x}_t)))\\
&\le p_t^*(1-F(p_t^*-v^*(\boldsymbol{x}_t)))-\dot{p}_t(1-F(\dot{p}_t-v^*(\boldsymbol{x}_t)))\\
&\le p_t^*(1-F(\dot{p}_t-v^*(\boldsymbol{x}_t)))-\dot{p}_t(1-F(\dot{p}_t-v^*(\boldsymbol{x}_t)))\\
&= (p_t^*-\dot{p}_t)(1-F(\dot{p}_t-v^*(\boldsymbol{x}_t)))\le p_t^*-\dot{p}_t,
\end{aligned}$$

where we used that $\dot{p}_t$ is the *left* neighbor of p_t^* and $1-F(\cdot)$ is nonincreasing and bounded by 1.

It remains to bound $p_t^*-\dot{p}_t$. Because the candidate prices $\{m_j+\hat{v}_k(\boldsymbol{x}_t)\}_{j=1}^{N_k}$ are equally spaced with grid spacing $(B+2\|\hat{v}_k\|_\infty)/N_k$, and m_j are midpoints, we have

$$m_1+\hat{v}_k(\boldsymbol{x}_t)\le\frac{B+2\|\hat{v}_k\|_\infty}{2N_k},\qquad m_{N_k}+\hat{v}_k(\boldsymbol{x}_t)\ge B-\frac{B+2\|\hat{v}_k\|_\infty}{2N_k}.$$

Three cases are possible:

(i) No candidate $\le p_t^*$. Then $\dot{p}_t=0$ and $p_t^*\le m_1+\hat{v}_k(\boldsymbol{x}_t)\le\frac{B+2\|\hat{v}_k\|_\infty}{2N_k}$.

(ii) Some candidate $\le p_t^*$ and the largest such candidate is not $m_{N_k}+\hat{v}_k(\boldsymbol{x}_t)$. Then the next candidate exceeds p_t^* and lies at distance at most $\frac{B+2\|\hat{v}_k\|_\infty}{N_k}$, hence $0\le p_t^*-\dot{p}_t\le\frac{B+2\|\hat{v}_k\|_\infty}{N_k}$.

(iii) The largest candidate $\le p_t^*$ is $m_{N_k}+\hat{v}_k(\boldsymbol{x}_t)$. By the midpoint bound above, $m_{N_k}+\hat{v}_k(\boldsymbol{x}_t)\ge B-\frac{B+2\|\hat{v}_k\|_\infty}{2N_k}$, so $0\le p_t^*-\dot{p}_t\le\frac{B+2\|\hat{v}_k\|_\infty}{2N_k}$.

In all cases,

$$0\le p_t^*-\dot{p}_t\le\frac{B+2\|\hat{v}_k\|_\infty}{N_k}.$$

Combining with the first part yields, for each round t,

$$p_t^*(1-F(p_t^*-v^*(\boldsymbol{x}_t)))-\tilde{p}_t^*(1-F(\tilde{p}_t^*-v^*(\boldsymbol{x}_t)))\le\frac{B+2\|\hat{v}_k\|_\infty}{N_k}\le\frac{3B}{N_k}.$$

Summing over the T_k rounds of episode k gives the stated episode-wise bound; summing over episodes yields the cumulative bound. □

Theorem 1. *Suppose $0<\delta<1/(2\lceil\log_2 T\rceil)$. Under Assumptions 1, 2 and 3, the regret of Algorithm 1 satisfies*

$$\mathrm{Reg}(T)=\tilde{\mathcal{O}}\Big(\rho_{\mathcal{V}}^{\frac{1}{3}}(\delta)T^{\frac{2}{3}}\Big)\quad\textit{with probability at least } 1-2\delta\lceil\log_2 T\rceil.$$

Proof. For episode k, recall that Γ_k is the high-probability event for UCB concentration. Let $\mathcal{E}_k$ denote the event that the offline oracle satisfies Assumption 4 with confidence level δ for the k-th call. By design, $\mathbb{P}(\Gamma_k)\ge1-\delta$ and $\mathbb{P}(\mathcal{E}_k)\ge1-\delta$. Applying the union bound, we have

$$\mathbb{P}\left(\bigcap_{k=1}^{\lceil\log_2 T\rceil}(\Gamma_k\cap\mathcal{E}_k)\right)\ge1-2\lceil\log_2 T\rceil\delta.$$

In what follows we work on the event $\bigcap_{k=1}^{\lceil\log_2 T\rceil}(\Gamma_k\cap\mathcal{E}_k)$.

Recall that $T_k^e = \lceil \ell_k^{\frac{2}{3}} \rho_{\mathcal{V}}^{\frac{1}{3}}(\delta) \rceil$, and whenever $T_k^e > \ell_k$ (which happens when $\ell_k < \rho_{\mathcal{V}}(\delta)$), we use a pure exploration episode, so these T_k^e are capped by ℓ_k. We know that the UCB phase begins no earlier than $k^* = \lceil \log_2(\rho_{\mathcal{V}}(\delta)) \rceil$. The regret during the exploration phase up to k^* is at most $B2^{k^*} \le B\rho_{\mathcal{V}}(\delta) + B$. Once the UCB phase begins, exactly T_k^e exploration rounds will be played, and since the price p_t is bounded by B, the regret during the exploration phase (with $k > k^*$) is

$$\sum_{k=k^*+1}^{\lceil \log_2 T \rceil} BT_k^e \le B\rho_{\mathcal{V}}^{\frac{1}{3}}(\delta) \sum_{k=1}^{\lceil \log_2 T \rceil} \ell_k^{\frac{2}{3}} \le 3B\rho_{\mathcal{V}}^{\frac{1}{3}}(\delta) T^{\frac{2}{3}},$$

where we used the fact that $\ell_k = 2^{k-1}$ and

$$\sum_{k=1}^{\lceil \log_2 T \rceil} \ell_k^{\frac{2}{3}} = \sum_{k=1}^{\lceil \log_2 T \rceil} 2^{\frac{2}{3}(k-1)} = \sum_{k=0}^{\lceil \log_2 T \rceil - 1} 2^{\frac{2}{3}k} = \frac{2^{\frac{2}{3}(\lceil \log_2 T \rceil)} - 1}{2^{\frac{2}{3}} - 1} \le \frac{(2T)^{2/3} - 1}{2^{\frac{2}{3}} - 1} \le 3T^{2/3}.$$

Combining this regret of exploration phase with Lemma 5 and Lemma 6, we have

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T} \mathcal{R}_t^1 + \sum_{t=1}^{T} \mathcal{R}_t^2 + \sum_{k=1}^{\lceil \log_2 T \rceil} BT_k^e \\
&\le \sum_{k=1}^{\lceil \log_2 T \rceil} \Big(16B\sqrt{2N_k T_k \ln(2S_k T_k N_k/\delta) \ln T_k} + 9BL\|\hat{v}_k - v^*\|_\infty T_k \ln T_k \\
&\qquad + 4BT_k^{\frac{1}{2}} + 64BN_k \ln(2S_k T_k N_k/\delta) + 3BT_k/N_k \Big) + B\rho_{\mathcal{V}}(\delta) + B + 3B\rho_{\mathcal{V}}^{\frac{1}{3}}(\delta) T^{\frac{2}{3}}.
\end{aligned}$$

Recall first that $T_k^e = \lceil \ell_k^{\frac{2}{3}} \rho_{\mathcal{V}}^{\frac{1}{3}}(\delta) \rceil$. By Assumption 4,

$$\|\hat{v}_k - v^*\|_\infty \le \sqrt{\frac{\rho_{\mathcal{V}}(\delta)}{T_k^e}} = \frac{\rho_{\mathcal{V}}^{1/3}(\delta)}{\ell_k^{1/3}} \quad \text{on} \quad \mathcal{E}_k.$$

Since $N_k = \lceil T_k^{\frac{1}{3}} / \ln^{\frac{1}{3}}(T_k/\delta) \rceil$, this implies that $\mathrm{Reg}(T) = \tilde{\mathcal{O}}(T^{\frac{2}{3}} \rho_{\mathcal{V}}^{\frac{1}{3}}(\delta))$ with probability at least $1 - 2\lceil \log_2(T) \rceil \delta$. □

B.3 Proofs for Lower Bounds in Section 4.2

B.3.1 Construction of Noise Distribution

We start with an infinitely differentiable function

$$u_0(x) = \begin{cases} \exp\left(-\dfrac{1}{x\left(\frac{1}{3} - x\right)} \right), & x \in \left(0, \frac{1}{3}\right), \\ 0, & \text{otherwise}, \end{cases}$$

which is nonnegative. Normalize u_0 via

$$u(x) = \Big(\int_0^{\frac{1}{3}} u_0(t)dt \Big)^{-1} \int_{-\infty}^{x} u_0(t)dt.$$

Then $u(x) = 0$ for $x \le 0$ and $u(x) = 1$ for $x \ge \frac{1}{3}$. For any positive integer l, the l-th derivative on $(0, \frac{1}{3})$ takes the form

$$\frac{\text{poly}(x)}{(x(\frac{1}{3}-x))^{2l-2}} \exp\left(-\frac{1}{x(\frac{1}{3}-x)}\right),$$

so for each $m \in \mathbb{N}$ there exists $L_m > 0$ with $\sup_x |u^{(m)}(x)|/m! \le L_m$, and $u^{(m)}(0) = u^{(m)}(\frac{1}{3}) = 0$ for $m \ge 1$. A Taylor expansion at $x = \frac{1}{3}$ with Lagrange remainder gives

$$\left|u(\tfrac{1}{3}) - u(x)\right| \le \frac{\sup_\xi |u^{(m)}(\xi)|}{m!} \left|x - \tfrac{1}{3}\right|^m \le L_m \left|x - \tfrac{1}{3}\right|^m.$$

We summarize the properties of $u(x)$ in the following proposition.

Proposition 1. *For the function u above:*

1. *$u(x)$ is nondecreasing;*
2. *$u^{(m)}(0) = u^{(m)}(\frac{1}{3}) = 0$ for all $m \ge 1$;*
3. *for every $m \ge 1$ and $x \in \mathbb{R}$, $|u(\frac{1}{3}) - u(x)| \le L_m |x - \frac{1}{3}|^m$.*

Next step is to construct a bump function

$$B(x) = \begin{cases} 0, & x < 0, \\ u(x), & 0 \le x \le \frac{1}{3}, \\ 1, & \frac{1}{3} < x < \frac{2}{3}, \\ u(1-x), & \frac{2}{3} \le x \le 1, \\ 0, & x > 1, \end{cases}$$

Since $u(x)$ is infinitely differentiable and its m-th order derivatives vanish at 0 and $\frac{1}{3}$, $B(x)$ is also infinitely differentiable. Furthermore, from Proposition 1, we have

$$|B(x) - B(\tfrac{1}{3})| \le L_m |x - \tfrac{1}{3}|^m. \tag{9}$$

For any $a < b$, let $B_{[a,b]}(x) = B(\frac{x-a}{b-a})$. Then (9) yields

$$\begin{aligned} B_{[a,b]}\left(a + \frac{b-a}{3}\right) - B_{[a,b]}(x) &= B\left(\frac{1}{3}\right) - B\left(\frac{x-a}{b-a}\right) \\ &\le L_m \left|\frac{1}{3} - \frac{x-a}{b-a}\right|^m = \frac{L_m}{(b-a)^m} \left|a + \frac{b-a}{3} - x\right|^m. \end{aligned}$$

Construct nested intervals $[0,1] = [a_0, b_0] \supset [a_1, b_1] \supset \cdots$ with lengths $w_k = b_k - a_k = 3^{-k!}$ for $k \ge 1$ (and $w_0 = 1$). We further partition $[a_{k-1} + \frac{w_{k-1}}{3}, b_{k-1} - \frac{w_{k-1}}{3}]$ into $Q_k = \frac{w_{k-1}}{3w_k}$ subintervals of length w_k and choose one of them as $[a_k, b_k]$. Note $B_{[a_k,b_k]}$ is constant on $[a_k + \frac{w_k}{3}, b_k - \frac{w_k}{3}]$, hence $B^{(\ell)}_{[a_k,b_k]}(x) = 0$ there for all $\ell \ge 1$ on that interval. Importantly, there exist infinitely many series of intervals that can be constructed in this manner. For each of these interval series, we define

$$f(x) = c_f \sum_{k=0}^{\infty} w_k^m B_{[a_k,b_k]}(x),$$

with $c_f > 0$ small (to be fixed). We list a few important properties of $f(x)$ below.

Proposition 2. *1.* $0 \le f(x) \le \frac{3}{2} c_f$ *for all* x.

2. There is a unique maximizer $x^* \in [0,1]$ *with* $x^* \in \cap_{k\ge 0}[a_k, b_k]$ *and* $f(x^*) = c_f \sum_{k=0}^{\infty} w_k^m$.

3. f *is unimodal: nondecreasing on* $[0, x^*]$ *and nonincreasing on* $[x^*, 1]$.

4. f *is* m*-times differentiable and* $\left|f^{(m)}(x)\right| \le c_f m! L_m$.

Proof. 1. Since $0 \le B_{[a_k,b_k]}(x) \le 1$, we have

$$0 \le f(x) \le \sum_{k=0}^{\infty} c_f w_k^m \le c_f \sum_{k=0}^{\infty} 3^{-k} \le \frac{3}{2} c_f < \infty.$$

2. Since $w_k = b_k - a_k = 3^{-k!} \to 0$, then there exists a unique x^* such that $x^* \in [a_k, b_k]$ for any k. For $B_{[a_k,b_k]}(x)$, it is nondecreasing on $(-\infty, x^*]$ and non-increasing on $[x^*, +\infty)$. Hence, x^* is the maximizer of $f(x)$. Since $x^* \in [a_{k+1}, b_{k+1}]$ and thus $B_{[a_k,b_k]}(x) = 1$, we immediately know that the maximum is $c_f \sum_{k=0}^{\infty} w_k^m \le \frac{3}{2} c_f$.

3. For any $k \ge 0$, $B_{[a_k,b_k]}(x)$ is nondecreasing in $[0, x^*]$ and non-increasing in $[x^*, 1]$ since $x^* \in [a_{k+1}, b_{k+1}] \subset [a_k, b_k]$. Thus $f(x)$, as the sum of these bump functions, is nondecreasing in $[0, x^*]$ and non-increasing in $[x^*, 1]$.

4. Since $B(x)$ is infinitely times differentiable, we have

$$\sum_{k=0}^{\infty} w_k^m B^{(m-1)}\left(\frac{x - a_k}{b_k - a_k}\right) \frac{1}{(b_k - a_k)^{m-1}} \le \sum_{k=0}^{\infty} w_k \le \sum_{k=0}^{\infty} 3^{-k} < \infty,$$

which shows that $c_f \sum_{k=0}^{\infty} w_k^m B_{[a_k,b_k]}^{(m-1)}(x)$ converges absolutely and uniformly. The above result implies $f(x)$ is $(m-1)$-th differentiable.

Notice that for any $x \in [0,1]$, there exists at most one k, such that $B_{[a_k,b_k]}^{(m)}(x) \ne 0$. Consider the case when $B_{[a_k,b_k]}^{(m)}(x) \ne 0$, and we know that

$$x \in [a_k, b_k] \subset \left[a_{k-1} + \frac{w_{k-1}}{3}, b_{k-1} - \frac{w_{k-1}}{3}\right] \subset \cdots [a_0, b_0] = [0,1].$$

Hence, we know that $B_{[a_j,b_j]}^{(m)}(x) = 0$ for $j = 0, 1, \cdots, k-1$. Moreover, we know that $x \notin [a_{k+1}, b_{k+1}]$; otherwise we have $B_{[a_k,b_k]}^{(m)}(x) = 0$. Therefore, we have $x \notin [a_j, b_j]$ for $j = k+1, k+2, \cdots$, which indicates $B_{[a_j,b_j]}^{(m)}(x) = 0$. Therefore, we have

$$c_f \max_{x\in[a_k,b_k],k\in\mathbb{N}} \left|w_k^m B_{[a_k,b_k]}^{(m)}(x)\right| \le c_f \max_{x\in[a_k,b_k],k\in\mathbb{N}} \left|B^{(m)}\left(\frac{x - a_k}{b_k - a_k}\right)\right| \le c_f m! L_m.$$

The above property implies $c_f \sum_{k=0}^{\infty} w_k^m B_{[a_k,b_k]}^{(m)}(x)$ also converges absolutely and uniformly. From the above property, we know that f is m-th differentiable.

□

We then rescale $f(x)$ to $[0,1]$ and define the function $g(x) = 1 - \frac{1}{1+f(x)}$. Notice that $g(x)$ is unimodal with the same unique maximizer x^* of $f(x)$. Furthermore,

$$|g(x^*) - g(x)| = \left|\frac{1}{1+f(x^*)} - \frac{1}{1+f(x)}\right| \le |f(x^*) - f(x)| \le \tilde{L}_m |x^* - x|^m.$$

We further define the function $F(x)$ as

$$F(x) = \begin{cases} 0, & \text{if } x < b, \\ 1 - \frac{b}{x} - \frac{1-b}{x} g\left(\frac{x-b}{1-b}\right), & \text{if } b \le x \le 1, \\ 2 - \frac{1+b}{x}, & \text{if } 1 < x \le 1+b, \\ 1, & \text{if } x > 1+b. \end{cases} \tag{10}$$

Let $c_f \in (0, 1/L_1)$ and $b = (1 + c_f L_1)/2 \in (0,1)$.

Proposition 3. *1. F is a right-continuous, nondecreasing CDF on $\mathbb{R}$.*

2. F is m-times differentiable on $(b,1)$.

3. The revenue $\mathsf{Rev}(x) = x(1-F(x))$ has a unique maximizer $x_r^* \in [b,1]$.

Proof. 1. It is easy to check that $F(-\infty) = 0, F(+\infty) = 1$ and $F(x)$ is continuous on $\mathbb{R}$. The remaining step is to show that $F(x)$ is nondecreasing. The monotonicity of $F(x)$ is clear on $(-\infty, b) \cup (1, +\infty]$. Then we consider the derivative of $F(x)$ on $[b,1]$. We have

$$F'(x) = \frac{b - x g'(\frac{x-b}{1-b}) + (1-b) g(\frac{x-b}{1-b})}{x^2}.$$

Since $|g'(x)| = \left|\frac{f'(x)}{(1+f(x))^2}\right| \le c_f L_1$, we have

$$b - x g'\left(\frac{x-b}{1-b}\right) \ge b - c_f L_1 = \frac{1 - c_f L_1}{2} > 0.$$

It indicates that $F'(x) > 0$ on $[b,1]$. Note that $F(b) \le F(x) \le F(1)$ for $x \in [b,1]$. Thus $F(x)$ is nondecreasing on $\mathbb{R}$.

2. It follows directly that $f(x)$ is m-th differentiable.

3. Simple calculation yields

$$\mathsf{Rev}(x) = \begin{cases} x, & \text{if } x \in [0, b), \\ b + (1-b) g\left(\frac{x-b}{1-b}\right), & \text{if } x \in [b, 1], \\ 1 + b - x, & \text{if } x \in (1, 1+b], \\ 0, & \text{if } x \in (1+b, \infty). \end{cases}$$

It is easy to check $\mathsf{Rev}(x) \ge b > \mathsf{Rev}(y)$ for any $x \in [b,1]$ and $y \in [0,+\infty) - [b,1]$. Since $g(x)$ has the same unique maximizer $x_g^* = x_f^*$ for f, we obtain that

$$x_r^* = b + (1-b) x_g^* \in [b,1]$$

is the unique maximizer for $\mathsf{Rev}(x)$.

□

B.3.2 Construction of Instances

Each interval sequence $\{[a_k, b_k]\}_{k\geq 0}$ yields a triplet (f, g, F) and thus an instance. With well-formulated groups of such instances, we will prove that no policy can perform well on all the instances in one group.

We work at a fixed level $k \geq 3$. In particular, we fix arbitrary level-i bumps for all $i \neq k$, and construct a group of instances that differs only in the level-k bump. Set

$$n_k \triangleq \left\lceil \frac{w_{k-1}}{k w_k^{2m+1}} \right\rceil, \quad Q_k = \frac{w_{k-1}}{3w_k}, \quad \text{and} \quad w_k = b_k - a_k = 3^{-k!}.$$

Let $I_1, \ldots, I_{Q_k}$ be the Q_k choices for $[a_k, b_k]$ at level k, each with width w_k. For each j, define the instance CDF F_j using f_j (and hence g_j) obtained by adding up all the fixed level-i bumps for $i \neq k$ and the level-k bump corresponding to I_j. Furthermore, let F_0 be the truncated reference CDF (with levels $< k$ only), defined using

$$f_0(x) = c_f \sum_{i=0}^{k-1} w_i^m B_{[a_i,b_i]}(x).$$

Recall that we set $c_f \in (0, 1/L_1)$ and $b = (1 + c_f L_1)/2 \in (0, 1)$. We restrict to prices $p_t \in [b, 1]$ since outside this range revenue is dominated by a price in $[b, 1]$. Let $u_{n_k} = (p_1, y_1, \ldots, p_{n_k}, y_{n_k})$ be the data generated under a policy π, and $\mathbb{P}_j$ (resp. $\mathbb{P}_0$) be the law under F_j (resp. F_0). Define the normalized price

$$q_t = (p_t - b)/(1 - b) \in [0, 1]$$

and the count

$$N_j = \sum_{t=1}^{n_k} \mathbb{I}\{q_t \in I_j\}.$$

B.3.3 Preliminary

The following lemmas will be useful.

Lemma 12 (Lemma 6, Luo et al. 2022)**.** *For Bernoulli distributions* $\mathrm{Ber}(p)$ *and* $\mathrm{Ber}(p+\varepsilon)$ *with* $\frac{1}{2} \leq p \leq p + \varepsilon \leq \frac{1}{2} + C$*, we have*

$$D_{\mathrm{KL}}(\mathrm{Ber}(p) \| \mathrm{Ber}(p+\varepsilon)) \leq \frac{4}{1 - 4C^2}\varepsilon^2.$$

Lemma 13 (Transportation inequality, a variant of Luo et al. 2022)**.** *Consider any function* h *on the sequence* u *that has a bounded value range* $[0, M]$*. Then for two probability measure* $\mathbb{P}_0$ *and* $\mathbb{P}_j$*,*

$$\mathbb{E}_{\mathbb{P}_j}[h(u)] - \mathbb{E}_{\mathbb{P}_0}[h(u)] \leq M\sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(\mathbb{P}_0 \| \mathbb{P}_j)}.$$

We remark that this is a variant (and direct corollary due to the symmetry of the total variation norm) of a result appeared in Luo et al. 2022, Appendix A.5, with $D_{\mathrm{KL}}(\mathbb{P}_0 \| \mathbb{P}_j)$ replacing $D_{\mathrm{KL}}(\mathbb{P}_j \| \mathbb{P}_0)$.

Lemma 14 (Instances difference bound)**.** *For any* $j \in [Q_k]$*,* $p \in [b, 1]$ *and* $q = (p - b)/(1 - b)$*:*

1. *If $q \notin I_j$, then $F_j(p) = F_0(p)$.*

2. *If $q \in I_j$, then*

$$0 \le (1 - F_j(p)) - (1 - F_0(p)) \le \tfrac{3}{2} c_f w_k^m, \quad \tfrac{1}{2} \le 1 - F_0(p) \le \tfrac{1}{1+c_f}, \quad \textit{and} \quad \tfrac{1}{2} \le 1 - F_j(p) \le \tfrac{1}{1+c_f}.$$

Proof. For p such that $q \notin I_j$, then (by the nesting) $q \notin [a_i, b_i]$ for every $i \ge k$. Hence $B_{[a_i,b_i]}(q) = 0$ for all $i \ge k$, and therefore

$$f_j(q) = c_f \sum_{i=0}^{k-1} w_i^m B_{[a_i,b_i]}(q) = f_0(q).$$

and hence $g_j(q) = g_0(q)$ and $F_j(p) = F_0(p)$.

Assume now $q \in I_j$. By construction,

$$f_j(q) - f_0(q) = c_f \sum_{i=k}^{\infty} w_i^m B_{[a_i,b_i]}(q) \ge 0.$$

and

$$\begin{aligned}
(1 - F_j(p)) - (1 - F_0(p)) &= \frac{1-b}{p}\Big(g_j(q) - g_0(q)\Big) = \frac{1-b}{p} \frac{f_j(q) - f_0(q)}{(1 + f_j(q))(1 + f_0(q))} \\
&\le \frac{1-b}{p}\Big(f_j(q) - f_0(q)\Big) = \frac{1-b}{p} c_f \sum_{i=k}^{\infty} w_i^m B_{[a_i,b_i]}(q) \\
&\le \frac{1-b}{b} c_f \sum_{i=k}^{\infty} w_i^m \le c_f \sum_{i=k}^{\infty} w_i^m,
\end{aligned}$$

since $p \ge b$ and $(1-b)/b \le 1$. To bound the tail $\sum_{i=k}^{\infty} w_i^m$, use

$$\frac{w_{k+r}^m}{w_k^m} = 3^{-m\left((k+r)!-k!\right)} \le 3^{-r} \qquad (r \ge 1),$$

hence

$$\sum_{i=k}^{\infty} w_i^m = w_k^m \sum_{r=0}^{\infty} \frac{w_{k+r}^m}{w_k^m} \le w_k^m \sum_{r=0}^{\infty} 3^{-r} = \tfrac{3}{2} w_k^m,$$

and therefore

$$0 \le (1 - F_j(p)) - (1 - F_0(p)) \le \tfrac{3}{2} c_f w_k^m.$$

It remains to bound the parameters $1 - F_0(p)$ and $1 - F_j(p)$ themselves. The arguements below works for both F_0 and F_j, hence we can drop the subscript. The lower bound is immediate from the generic relationship

$$1 - F(p) = \frac{b + (1-b)g(q)}{p} \ge \frac{b}{p} \ge b \ge \tfrac{1}{2},$$

since $g \ge 0$, $p \le 1$, and $b = \frac{1+c_f L_1}{2} \ge \frac{1}{2}$.

For the *upper bound*, note first that F is nondecreasing on $[b, 1]$ (see the monotonicity proof in the construction), so $1 - F$ is nonincreasing. Because $q \in I_j \subset [1/3, 2/3]$, we have $p = b + (1-b)q \ge b + (1-b)/3$. Thus

$$1 - F(p) \le 1 - F\Big(b + (1-b)\tfrac{1}{3}\Big) = \frac{b + (1-b)g(\frac{1}{3})}{b + (1-b)\frac{1}{3}}.$$

At $x = \frac{1}{3}$, the bump sum satisfies $f(\frac{1}{3}) = c_f$, hence

$$g(\tfrac{1}{3}) = \frac{f(\frac{1}{3})}{1 + f(\frac{1}{3})} = \frac{c_f}{1 + c_f}.$$

Combining the last two displays and writing $b = \frac{1+c_f L_1}{2}$ gives

$$1 - F(p) \le \frac{b + (1-b)\dfrac{c_f}{1+c_f}}{b + (1-b)\frac{1}{3}} = \frac{3(b + c_f)}{(2b+1)(c_f+1)}.$$

We further choose $c_f < \frac{1}{L_1+6}$ and recall that $b = \frac{1+c_f L_1}{2}$. Hence we have $3(b + c_f) < 2b + 1$ and

$$1 - F(p) \le \frac{1}{1 + c_f} < 1.$$

This completes the proof. □

B.3.4 Information Bounds

Lemma 15 (Per-round KL bound)**.** *For any $j \in [Q_k]$ and $p \in [b, 1]$ with $q = (p - b)/(1 - b)$,*

$$D_{\mathrm{KL}}(\mathrm{Ber}(1 - F_0(p)) \| \mathrm{Ber}(1 - F_j(p))) \le \frac{1}{300} w_k^{2m} \mathbb{I}\{q \in I_j\}.$$

Proof. If $q \notin I_j$, the Bernoulli parameters coincide by Lemma 14(1). If $q \in I_j$, then the parameters lie in $[\frac{1}{2}, \frac{1}{1+c_f}]$ and their gap is $\le \frac{3}{2} c_f w_k^m$ by Lemma 14(2). Applying Lemma 12, we obtain

$$\begin{aligned} D_{\mathrm{KL}}(\mathrm{Ber}(1 - F_0(p_t)) \| \mathrm{Ber}(1 - F_j(p_t))) &\le \frac{4}{1 - 4(\frac{1}{1+c_f} - \frac{1}{2})^2} ((1 - F_j(p_t)) - (1 - F_0(p_t)))^2 \\ &= \frac{9}{4} c_f (1 + c_f)^2 w_k^{2m} \\ &\le \frac{1}{300} w_k^{2m}. \end{aligned}$$

The last inequality holds as we can choose positive c_f such that $0 < c_f < \min\{10^{-4}, \frac{1}{L_1+6}\}$. □

By chain rule of KL, summing the per-round KLs from Lemma 15 and taking the corresponding expectation, we immediately obtain the bound for pathwise KL accumulation.

Lemma 16 (Pathwise KL accumulation)**.** *For any $j \in [Q_k]$, we have*

$$D_{\mathrm{KL}}(\mathbb{P}_0 \| \mathbb{P}_j) \le \tfrac{1}{300} w_k^{2m} \mathbb{E}_{\mathbb{P}_0}[N_j].$$

Lemma 17 (Bounding expectation of N_j)**.** *For all $k \ge 3$,*

$$\frac{1}{Q_k} \sum_{j=1}^{Q_k} \mathbb{E}_{\mathbb{P}_j}[N_j] \le \frac{1}{5} n_k.$$

In particular, there exists $j^\star \in [Q_k]$ with $\mathbb{E}_{\mathbb{P}_{j^\star}}[N_{j^\star}] \le \frac{1}{5} n_k$.

Proof. Note that N_j is a function of the price and response sequence u_{n_k} and is bounded by n_k. Apply Lemma 13 (with $h = N_j$ and $M = n_k$) and Lemma 16 to get

$$\mathbb{E}_{\mathbb{P}_j}[N_j] - \mathbb{E}_{\mathbb{P}_0}[N_j] \le n_k\sqrt{\tfrac{1}{2}D_{\mathrm{KL}}(\mathbb{P}_0\|\mathbb{P}_j)} \le n_k w_k^m\sqrt{\tfrac{1}{600}\mathbb{E}_{\mathbb{P}_0}[N_j]}.$$

For $k \ge 3$, we sum over $j \in [Q_k]$ take the average to obtain

$$\begin{aligned}
\frac{1}{Q_k}\sum_{j=1}^{Q_k}\mathbb{E}_{\mathbb{P}_j}[N_j] &\le \frac{1}{Q_k}\sum_{j=1}^{Q_k}\mathbb{E}_{\mathbb{P}_0}[N_j] + \frac{n_k w_k^m}{20Q_k}\sum_{j=1}^{Q_k}\sqrt{\mathbb{E}_{\mathbb{P}_0}[N_j]} = \frac{n_k}{Q_k} + \frac{n_k w_k^m}{20Q_k}\sum_{j=1}^{Q_k}\sqrt{\mathbb{E}_{\mathbb{P}_0}[N_j]} \\
&\le \frac{n_k}{Q_k} + \frac{n_k w_k^m}{20Q_k}\sqrt{Q_k\sum_{j=1}^{Q_k}\mathbb{E}_{\mathbb{P}_0}[N_j]} = \frac{n_k}{Q_k} + \frac{n_k w_k^m}{20Q_k}\sqrt{Q_k n_k} \\
&= n_k\left(\frac{1}{Q_k} + \frac{w_k^m}{20}\sqrt{\frac{2w_{k-1}}{w_k^{2m+1}Q_k}}\right) \\
&\le n_k\left(\frac{1}{27} + \frac{\sqrt{6}}{20}\right) \\
&\le \frac{1}{5}n_k.
\end{aligned}$$

Therefore, there exists some $j^\star \in [Q_k]$ such that $\mathbb{E}_{\mathbb{P}_{j^\star}}[N_{j^\star}] \le \frac{1}{5}n_k$. □

B.3.5 Completing the Lower Bound Proof

Lemma 18 (Revenue gap)**.** *Let $p_j^\star$ maximize Rev_j and set $q_j^\star = (p_j^\star - b)/(1-b) \in I_j$. There exists $\tilde{C} \triangleq \frac{1-c_f L_1}{2}\frac{c_f}{(1+\frac{3}{2}c_f)^2} > 0$ such that for any $p \in [b,1]$ with $q = (p-b)/(1-b) \notin I_j$,*

$$\mathsf{Rev}_j(p_j^\star) - \mathsf{Rev}_j(p) \ge \tilde{C}w_k^m.$$

Proof. For any $p \in [b,1]$ such that $q = \frac{p-b}{1-b} \notin I_j$, $B_{[a_i,b_i]}(q)$ is equal to zero for $i \ge k$ as $I_j \supset [a_{k+1}, b_{k+1}] \supset \cdots$. Recall that

$$f_j(x) = c_f\sum_{i=0}^{k-1} w_i^{2m}B_{[a_i,b_i]}(x) + c_f w_i^{2m}B_{[a_k,b_k]}(x) + c_f\sum_{i=k+1}^{\infty} w_i^{2m}B_{[a_i,b_i]}(x).$$

Hence, $f_j(q_j^\star) - f_j(q) \ge c_f w_k^m$, and therefore $g_j(q_j^\star) - g_j(q) \ge \frac{c_f}{(1+\frac{3}{2}c_f)^2}w_k^m$. Since $\mathsf{Rev}_j(p) = b + (1-b)g_j(q)$ on $[b,1]$ and $1-b = (1-c_f L_1)/2$,

$$\mathsf{Rev}_j(p_j^\star) - \mathsf{Rev}_j(p) \ge \frac{1-c_f L_1}{2}\cdot\frac{c_f}{(1+\frac{3}{2}c_f)^2}w_k^m = \tilde{C}w_k^m.$$

□

Lemma 19 (Regret at horizon n_k)**.** *For the $j^\star$ in Lemma 17,*

$$\mathbb{E}_{\mathbb{P}_{j^\star}}[\mathrm{Reg}(n_k)] \ge 0.8\tilde{C}w_k^m n_k.$$

Proof. Split rounds by $\{q_t \in I_{j^\star}\}$. On $\{q_t \notin I_{j^\star}\}$, Lemma 18 yields a per-round gap $\geq \tilde{C} w_k^m$; otherwise the gap is ≥ 0. Thus

$$\mathbb{E}_{\mathbb{P}_{j^\star}}[\mathrm{Reg}(n_k)] \geq \tilde{C} w_k^m \mathbb{E}_{\mathbb{P}_{j^\star}}[n_k - N_{j^\star}] \geq \tilde{C} w_k^m (n_k - \tfrac{1}{5} n_k) = 0.8 \tilde{C} w_k^m n_k.$$

□

We are finally ready to establish the lower bound.

Proof of Theorem 2. With $n_k = \lceil w_{k-1}/(k w_k^{2m+1}) \rceil$ and $w_k = 3^{-k!}$, for all sufficiently large k,

$$w_k^m n_k \geq n_k^{\frac{m+1-1/k}{2m+1-1/k}} \geq n_k^{\frac{m+1}{2m+1} - \frac{m}{k(2m+1)^2}}. \tag{11}$$

By Lemma 19 and (11), for large k,

$$\mathbb{E}_{\mathbb{P}_{j^\star}}[\mathrm{Reg}(n_k)] \geq 0.8 \tilde{C} n_k^{\frac{m+1}{2m+1} - \frac{m}{k(2m+1)^2}}.$$

If a policy achieved $\mathcal{O}(T^{\frac{m+1}{2m+1} - \zeta})$ regret for some $\zeta > 0$ uniformly over F, then choosing k large enough so that $\frac{m}{k(2m+1)^2} \leq \zeta/2$ would contradict the bound above at horizon $T = n_k$:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_{j^\star}}[\mathrm{Reg}_\pi(n_k)] &\leq C_\pi n_k^{\frac{m+1}{2m+1} - \zeta} < C_\pi n_k^{-\frac{m}{k(2m+1)^2}} n_k^{\frac{1+m}{2m+1} - \frac{m}{k(2m+1)^2}} \\
&< 0.8 \tilde{C} n_k^{\frac{1+m}{2m+1} - \frac{m}{k(2m+1)^2}} < \mathbb{E}_{\mathbb{P}_{j^\star}}[\mathrm{Reg}_\pi(n_k)].
\end{aligned}$$

Hence no policy can guarantee $\mathcal{O}(T^{\frac{m+1}{2m+1} - \zeta})$ regret for any $\zeta > 0$. □

B.4 Proofs for Section 5

Corollary 2. *Suppose $0 < \delta < 1/(2\lceil \log_2(T) \rceil)$. Under Assumptions 1, 2, 3 and 6, the regret of Algorithm 1 with $T_k^e = \rho_{\mathcal{V}}^{\frac{1}{2+\alpha}}(\delta) \ell_k^{\frac{2}{2+\alpha}}$ satisfies*

$$\mathrm{Reg}(T) = \tilde{\mathcal{O}}\big((\rho_{\mathcal{V}}^{\frac{1}{3}}(\delta) T^{\frac{2}{3}}) \vee (\rho_{\mathcal{V}}^{\frac{1}{2+\alpha}}(\delta) T^{\frac{2}{2+\alpha}})\big) \quad \textit{with probability at least } 1 - 2\lceil \log_2(T) \rceil \delta.$$

Proof. First, we analyze the regret of k-th episode on the event Γ_k and Assumption 6. Recall definitions $T_k^e = \rho_{\mathcal{V}}^{\frac{1}{2+\alpha}}(\delta) \ell_k^{\frac{2}{2+\alpha}}$, $T_k = \ell_k - T_k^e$ and $N_k = \left\lceil T_k^{\frac{1}{3}} / \rho_{\mathcal{V}}^{\frac{1}{3}}(\delta) \right\rceil$ in Algorithm 1. From Lemma 5 and Lemma 6, we know that

1. the regret from learning F scales as $\tilde{\mathcal{O}}(\sqrt{N_k T_k}) = \tilde{\mathcal{O}}(T^{\frac{2}{3}} / \rho_{\mathcal{V}}^{\frac{1}{6}}(\delta))$;
2. the regret from discretization error scales as $\tilde{\mathcal{O}}(T_k / N_k) = \tilde{\mathcal{O}}(\rho_{\mathcal{V}}^{\frac{1}{3}}(\delta) T^{\frac{2}{3}})$, which dominates the regret from learning F (without loss of generality we assume $\rho_{\mathcal{V}}(\delta) > 1$);
3. the length of the exploration phase scales as $\mathcal{O}(T_k^e) = \mathcal{O}(\rho_{\mathcal{V}}^{\frac{1}{2+\alpha}}(\delta) \ell_k^{\frac{2}{2+\alpha}})$; and
4. regret from estimation error of v^* scales with

$$\tilde{\mathcal{O}}(\|\hat{v}_k - v^*\|_\infty T_k) = \tilde{\mathcal{O}}\Big(\sqrt{\rho_{\mathcal{V}}(\delta) / (T_k^e)^\alpha} T_k\Big) = \tilde{\mathcal{O}}(\rho_{\mathcal{V}}^{\frac{1}{2+\alpha}}(\delta) \ell_k^{\frac{2}{2+\alpha}}).$$

Combing the all terms and applying the union bounds yield the desired result. □

Corollary 3. *Suppose* $0 < \delta < 1/(2\lceil \log_2 T \rceil)$. *Under Assumptions 1, 2, 5 and 7, the regret of Algorithm 3 satisfies*

$$\mathrm{Reg}(T) = \tilde{\mathcal{O}}(T^{\frac{3}{5}} \vee \rho_{\mathcal{V}}^{\frac{1}{2}}(\delta) T^{1-\frac{\alpha}{2}}) \quad \textit{with probability at least } 1 - 2\lceil \log_2(T) \rceil \delta.$$

Proof. Since we invoke a classification oracle, no separate exploration phase is needed: we use samples of size $T_{k-1} = \frac{1}{2} T_k$ from the previous episode. By Assumption 7, this guarantees $\|\hat{v}_k - v^*\|_\infty = \mathcal{O}(\rho_{\mathcal{V}}^{\frac{1}{2}}(\delta) T_k^{-\frac{\alpha}{2}})$ with probability at least $1 - \delta$. In what follows, we focus on the asymptotic order of the regret, omitting constant factors and logarithmic terms for brevity.

From Lemma 5, the learning regret is bounded by

$$\sum_{k=1}^{\lceil \log_2 T \rceil} \Big[16B \sqrt{2 N_k T_k \ln(2 S_k T_k N_k / \delta) \ln T_k} + 9BL \|\hat{v}_k - v^*\|_\infty T_k \ln T_k + 4B T_k^{\frac{1}{2}} + 64 B N_k \ln(2 S_k T_k N_k / \delta) \Big] \quad \text{with probability at least } 1 - \lceil \log_2 T \rceil \delta.$$

Substituting $N_k = \lceil T_k^{\frac{1}{5}} \rceil$ and the error bound $\|\hat{v}_k - v^*\|_\infty = \mathcal{O}(\rho_{\mathcal{V}}^{\frac{1}{2}}(\delta) T_k^{-\frac{\alpha}{2}})$, the dominant term in the summation simplifies to $\tilde{\mathcal{O}}(T^{\frac{3}{5}} \vee \rho_{\mathcal{V}}^{\frac{1}{2}}(\delta) T^{1-\frac{\alpha}{2}})$, with other terms (e.g., $BL\sqrt{T}$, $T^{\frac{1}{5}}$) being asymptotically negligible. The discretization regret is bounded by $\tilde{\mathcal{O}}\left(\sum_{k=1}^{\lceil \log_2 T \rceil} T_k / N_k^2\right) = \tilde{\mathcal{O}}(T^{\frac{3}{5}})$, due to Assumption 5. Combining the learning regret and discretization regret, we thus obtain the desired bound $\mathrm{Reg}(T) = \tilde{\mathcal{O}}(T^{\frac{3}{5}} \vee \rho_{\mathcal{V}}^{\frac{1}{2}}(\delta) T^{1-\frac{\alpha}{2}})$. □

Assumption 8 (Differentiability)**.** *The function F is twice continuously differentiable.*

Assumption 9 (Concavity)**.** *The function F and $1 - F$ is log-concave.*

Corollary 4. *Assume that the noise distribution F is twice continuously differentiable, and that both F and $1 - F$ are log-concave. Under Assumptions 1 and 4, the regret of Algorithm 4 satisfies*

$$\mathrm{Reg}(T) = \mathcal{O}(\rho_{\mathcal{V}}(\delta) \ln T) \quad \textit{with probability at least } 1 - \lceil \log_2 T \rceil \delta.$$

Proof. Fix an episode $k \geq 2$. By the first-order optimality condition, the optimal price in episode k is $p_t = g(\hat{v}_k(\boldsymbol{x}_t))$, where $g(v) = v + \phi^{-1}(-v)$ and $\phi(v) = v - \frac{1-F(v)}{F'(v)}$. Since $1 - F$ is log-concave, the hazard $h(v) = \frac{F'(v)}{1-F(v)}$ is increasing, hence $\phi'(v) = 1 + \frac{h'(v)}{h(v)^2} \geq 1$, so ϕ is strictly increasing and g is 1-Lipschitz: $|g(v) - g(w)| \leq |v - w|$.

Given $\hat{v}_k$ (fit on the previous episode), the random variables $\{(\boldsymbol{x}_t, p_t, y_t)\}$ in episode k are i.i.d. because p_t depends only on $\boldsymbol{x}_t$ and the covariates are i.i.d. We temporarily abuse the revenue function notation and write $\mathsf{Rev}_q(p) = p(1 - F(p - q))$. Let $q_t = v^*(\boldsymbol{x}_t)$ and $p_t^* = g(q_t)$ denote the episode-k optimal price. By Taylor's theorem around p_t^*,

$$\mathsf{Rev}_t(p_t^*) - \mathsf{Rev}_t(p_t) = -\tfrac{1}{2} \mathsf{Rev}''_{q_t}(\chi_t)(p_t - p_t^*)^2 \leq \tfrac{1}{2} C |p_t - p_t^*|^2,$$

for some χ_t between p_t and p_t^*, where

$$C \triangleq \sup_{q,p\in[0,B]} |\mathsf{Rev}_q''(p)| < \infty$$

is finite because $F \in C^2$ and $(p-q)$ ranges over a compact set (by boundedness of p and v^*). Since g is 1-Lipschitz, $|p_t - p_t^*| \le \|\hat{v}_k - v^*\|_\infty$ and thus

$$\mathsf{Rev}_t(p_t^*) - \mathsf{Rev}_t(p_t) \le \tfrac{1}{2} C \|\hat{v}_k - v^*\|_\infty^2.$$

Let $\ell_{k-1} = 2^{k-2}$ be the size of the previous episode used to fit $\hat{v}_k$. Under the known-F offline regression oracle, for $\delta > 0$ we have with probability at least $1-\delta$,

$$\|\hat{v}_k - v^*\|_\infty \le \sqrt{\rho_{\mathcal{V}}(\delta)/\ell_{k-1}}.$$

Therefore, on this event,

$$\mathsf{Rev}_t(p_t^*) - \mathsf{Rev}_t(p_t) \le \tfrac{1}{2} C \frac{\rho_{\mathcal{V}}(\delta)}{\ell_{k-1}} = \tfrac{1}{2} C \frac{\rho_{\mathcal{V}}(\delta)}{2^{k-2}}.$$

Summing over the 2^{k-1} rounds of episode k gives

$$\sum_{t=2^{k-1}}^{2^k} \left(\mathsf{Rev}_t(p_t^*) - \mathsf{Rev}_t(p_t)\right) \le C\rho_{\mathcal{V}}(\delta).$$

Apply the union bound over episodes $k = 2, \dots, \lceil \log_2 T \rceil$ to get that, with probability at least $1 - \lceil \log_2 T \rceil \delta$,

$$\mathrm{Reg}(T) \le B + \sum_{k=2}^{\lceil \log_2 T \rceil} C\rho_{\mathcal{V}}(\delta) = \mathcal{O}\big(\rho_{\mathcal{V}}(\delta) \log T\big).$$

□

Corollary 5. *Under Assumptions 1, 3 and 4, the regret of Algorithm 4 satisfies*

$$\mathrm{Reg}(T) = \mathcal{O}\big(\sqrt{\rho_{\mathcal{V}}(\delta) T \ln T}\big) \quad \textit{with probability at least } 1 - \lceil \log_2 T \rceil \delta.$$

Proof. We first consider one-step regret:

$$\begin{aligned}
&\mathsf{Rev}_t(p_t^*) - \mathsf{Rev}_t(p_t) \\
&= p_t^*(1 - F(p_t^* - v^*(\boldsymbol{x}_t))) - p_t(1 - F(p_t - v^*(\boldsymbol{x}_t))) \\
&= p_t^*(1 - F(p_t^* - v^*(\boldsymbol{x}_t))) - p_t^*(1 - F(p_t^* - \hat{v}_k(\boldsymbol{x}_t))) + p_t^*(1 - F(p_t^* - \hat{v}_k(\boldsymbol{x}_t))) \\
&\qquad - p_t(1 - F(p_t - \hat{v}_k(\boldsymbol{x}_t))) + p_t(1 - F(p_t - \hat{v}_k(\boldsymbol{x}_t))) - p_t(1 - F(p_t - v^*(\boldsymbol{x}_t))) \\
&\le p_t^*(1 - F(p_t^* - v^*(\boldsymbol{x}_t))) - p_t^*(1 - F(p_t^* - \hat{v}_k(\boldsymbol{x}_t))) + p_t(1 - F(p_t - \hat{v}_k(\boldsymbol{x}_t))) \\
&\qquad - p_t(1 - F(p_t - v^*(\boldsymbol{x}_t))) \\
&\le 2BL|v^*(\boldsymbol{x}_t) - \hat{v}_k(\boldsymbol{x}_t)|.
\end{aligned}$$

The first inequality is due to the optimality of p_t with respect to $\hat{v}_k$. Therefore, we have

$$\mathsf{Rev}_t(p_t^*) - \mathsf{Rev}_t(p_t) \le 2BL|v^*(\boldsymbol{x}_t) - \hat{v}_k(\boldsymbol{x}_t)| \le 2BL\|v^* - \hat{v}_k\|_\infty.$$

Recall the estimation guarantee of $\|\hat{v}_k - v^*\|_\infty \le \sqrt{\rho_{\mathcal{V}}(\delta)/\ell_{k-1}}$ with probability at least $1-\delta$ for episode k. Applying the union bound over episodes $k = 1, \dots, \lceil \log_2 T \rceil$, and summing over all rounds in all episodes and plugging in the estimation guarantee, we obtain the desired result. □

Algorithm Algorithm 5 requires an offline regression oracle for i.i.d. samples $\{(\boldsymbol{x}_t, v_t)\}$ that satisfy the moment condition $\mathbb{E}[v_t \mid \boldsymbol{x}_t] = v^*(\boldsymbol{x}_t)$. We state this assumption formally below.

Assumption 10 (Adjusted Offline Regression Oracle)**.** *Under realizability Assumption 1, let $\{(\boldsymbol{x}_t, v_t)\}_{t\in[n]}$ be i.i.d. samples from a fixed but unknown distribution, satisfying $\mathbb{E}[v_t \mid \boldsymbol{x}_t] = v^*(\boldsymbol{x}_t)$. Given these samples and any confidence level $\delta > 0$, an offline regression oracle returns a predictor $\hat{v} \in \mathcal{V}$ such that*

$$\|\hat{v} - v^*\|_\infty \le \sqrt{\rho_{\mathcal{V}}(\delta)/n} \quad \textit{with probability at least } 1-\delta.$$

Corollary 6. *Suppose $0 < \delta < 1/(2\lceil \log_2 T \rceil)$. Under Assumptions 1, 2, 10 and 5, the regret of Algorithm 5 satisfies*

$$\mathrm{Reg}(T) = \tilde{\mathcal{O}}(T^{\frac{3}{5}} \vee \rho_{\mathcal{V}}^{\frac{1}{2}}(\delta) T^{\frac{1}{2}}) \quad \textit{with probability at least } 1 - 2\lceil \log_2(T) \rceil \delta.$$

Proof. Since v_t is directly observable, exploration phase is unnecessary so no additional regret arises. The sample size for the adjusted offline regression oracle at the episode k is $T_{k-1} = \frac{1}{2} T_k$. This ensures the estimation error bound $\|\hat{v}_k - v^*\|_\infty = \mathcal{O}(\rho_{\mathcal{V}}^{\frac{1}{2}}(\delta) T_k^{-\frac{1}{2}})$. In what follows, we focus on the asymptotic order of the regret, omitting constant factors and logarithmic terms for brevity.

From Lemma 5, the learning regret is bounded by

$$\sum_{k=1}^{\lceil \log_2 T \rceil} \Big[16B\sqrt{2N_k T_k \ln(2S_k T_k N_k/\delta) \ln T_k} + 9BL\|\hat{v}_k - v^*\|_\infty T_k \ln T_k + 4BT_k^{\frac{1}{2}} + 64BN_k \ln(2S_k T_k N_k/\delta) \Big] \quad \text{with probability at least } 1 - \lceil \log_2 T \rceil \delta.$$

Substituting $N_k = \lceil T_k^{\frac{1}{5}} \rceil$ and the error bound $\|\hat{v}_k - v^*\|_\infty = \mathcal{O}(\rho_{\mathcal{V}}^{\frac{1}{2}}(\delta) T_k^{-\frac{1}{2}})$, the dominant term in the summation simplifies to $\tilde{\mathcal{O}}(T^{\frac{3}{5}})$, with other terms (e.g., $BL\sqrt{T}$, $T^{\frac{1}{5}}$) being asymptotically negligible. The discretization regret is bounded by $\tilde{\mathcal{O}}\left(\sum_{k=1}^{\lceil \log_2 T \rceil} T_k / N_k^2\right) = \tilde{\mathcal{O}}(T^{\frac{3}{5}})$, due to Assumption 5. Combining the learning regret and discretization regret, we thus obtain the desired bound $\mathrm{Reg}(T) = \tilde{\mathcal{O}}(T^{\frac{3}{5}} \vee \rho_{\mathcal{V}}^{\frac{1}{2}}(\delta) T^{\frac{1}{2}})$. □