NONPARAMETRIC CONTEXTUAL PRICING AND INVENTORY LEARNING UNDER CENSORED DEMAND

Anonymous authors
Paper under double-blind review

ABSTRACT

In online retailing, when a product sells out, a retailer often sees only the units sold, not how many customers would have bought it had inventory been available. However, the inventory level determines how much demand is revealed, and this information can influence subsequent decisions and future profits. We study an online selling problem in which, in each round, the seller observes a market context and then makes pricing and stocking decisions based on censored sales data from previous rounds. The challenge is to learn a context-dependent pricing and stocking policy without a parametric demand model or fully observed profit, since lost-sales costs depend on unobserved unmet demand. To overcome this difficulty, we propose a Surrogate-Profit UCB (SP-UCB) algorithm that turns each incomplete sales record into a reliable guide for both inventory and price decisions, using data from past rounds with similar market conditions. This design allows us to learn while serving customers, without a separate exploration phase or the need to recover all demand hidden by stockouts. We prove the minimax optimality of the proposed algorithm, with strictly faster rates when expected profit varies more smoothly with price. Comprehensive numerical experiments have been conducted to confirm the effectiveness of the proposed algorithm.

1 INTRODUCTION

Online retailers often need to choose both a price and an inventory level before demand is known. Flash-sale and fashion retailers, for example, sell short-lived assortments and face changing seasons and local markets. They use limited sales records to forecast demand and set prices (Ferreira et al., 2016; Caro & Gallien, 2012). Price and inventory are closely linked: a lower price can increase demand and the risk of a stockout, while a cautious inventory level can reduce leftover costs but also hide how much demand the price would have generated.

The data generated by these operations are censored. When a product sells out, observed sales equal the available inventory, leaving unmet demand unknown (Musalem et al., 2010; Conlon & Mortimer, 2013). Because lost-sales costs depend on this unobserved demand, the retailer cannot fully observe realized profit. Learning from sales therefore requires a way to compare the profitability of pricing and stocking decisions despite the missing demand (Besbes & Muharremoglu, 2013; Chen et al., 2021; 2024).

The problem becomes harder when decisions must adapt to context. The same price may work well in one season or local market and poorly in another. Pooling all past sales would ignore these differences, while learning separately in every context would leave too little data. Even at the same price and in similar contexts, different inventory levels can produce different observed sales. The seller must therefore share information across similar contexts while accounting for the censoring caused by past inventory decisions.

Existing online studies of censored pricing and inventory learning focus on a fixed demand environment (Chen et al., 2021; 2024), while contextual approaches often use historical data collected before deployment (Tang et al., 2025; Gundem & Qi, 2025). Recent work also studies online learning under contextual and censored demand, using a finite collection of basis functions and an initial high-inventory exploration phase (Han et al., 2026a). Our focus is on learning under a nonparametric demand model, without requiring inventories high enough to reveal all demand. This raises a natural question:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

How can an online seller use censored sales to learn which prices and inventory levels are best for each context when realized profit is not fully observed, and what is the best achievable learning rate?

We answer this question through three contributions.

1. **An observable learning signal under sales censoring.** With sales-only feedback, the seller cannot tell whether observed sales reflect true demand or the inventory cap. We construct an observable score from sales and leftover inventory. At a fixed price, this score has the same best inventory choice as true profit. For comparing prices, however, the score has a price-dependent bias. We remove that bias using a low-demand quantile that the algorithm makes visible. The resulting signal contains the information needed for both stocking and pricing decisions.
2. **Context-aware learning with censored sales.** We develop the Surrogate-Profit UCB (SP-UCB) algorithm for joint pricing and stocking. Each inventory decision makes the calibration quantile observable, while context-weighted sales records provide the information used to evaluate prices. All learning takes place during deployment, using the sales information generated by regular operations.
3. **Minimax-optimal regret rates.** We prove matching upper and lower bounds for the proposed method. Under Lipschitz price values, the minimax regret is $\widetilde{\Theta}\left(T^{\frac{d+2\alpha}{d+3\alpha}}\right)$, while under twice-smooth price values it becomes $\widetilde{\Theta}\left(T^{\frac{2d+3\alpha}{2d+5\alpha}}\right)$. These rates quantify two sources of difficulty: a higher context dimension makes local learning harder, while a smoother expected-profit curve over price allows a coarser price grid and faster learning.

2 Related Work

Censored inventory and joint pricing–inventory learning. Learning inventory decisions from censored sales has a long history. Existing methods use sales and stockout information to learn stocking policies without observing unmet demand (Huh & Rusmevichientong, 2009; Huh et al., 2011; Besbes & Muharremoglu, 2013), with subsequent work incorporating contextual features (Ding et al., 2024). Joint pricing introduces the additional need to compare profits across prices. In a noncontextual setting, Chen et al. (2021) construct a spline approximation of the demand–price function, while Chen et al. (2024) estimate profit differences between prices through an inversion of empirical distributions. Our approach first constructs an observable sales score whose expectation preserves inventory profit gaps at a fixed price. Under a residual law shared across prices within each context, a visible low quantile shifts with mean demand and can therefore remove the score's price-dependent bias. The resulting calibrated target preserves profit gaps across prices. We estimate this target locally across contexts, allowing the residual distribution to vary with context without recovering the unobserved demand tail.

Contextual pricing and bandit learning. Contextual pricing methods use customer or product features to learn how demand responds to prices, often through structured demand models (Cohen et al., 2020; Javanmard & Nazerzadeh, 2019; Ban & Keskin, 2021). Nonparametric pricing exploits smoothness across prices or contexts (Chen & Gallego, 2021; Wang et al., 2021). Contextual bandit methods exploit similarity (Slivkins, 2014), while continuous-action methods also study smoothed policy benchmarks (Krishnamurthy et al., 2020). For offline-to-online linear contextual bandits, Han et al. (2026b) use directional bias bounds to guide the reuse of historical data. Related work studies dynamic pricing under inventory constraints (Chen & Gallego, 2022) and revenue-based financing (Lin et al., 2026). In our setting, the inventory chosen in each period also determines how much demand becomes visible. When a stockout occurs, lost-sales costs are unobserved, so realized profit cannot be used directly as a learning signal. Optimistic price selection therefore requires both a calibrated signal and inventory decisions that make it estimable. Inventory estimates need only be accurate enough to support price comparison: near the optimal stocking level, profit loss is quadratic in inventory error. SP-UCB exploits this property to learn inventories on a coarser context partition, reducing the cost of inventory learning while controlling the error in price evaluation.

Contextual joint pricing and inventory learning. Historical-data methods address feature-based inventory control (Ban & Rudin, 2019) and joint pricing–inventory optimization (Qin et al., 2022).

With censored sales, Bu et al. (2023) learn prices for a given inventory, Tang et al. (2025) learn feature-based pricing rules, and Gundem & Qi (2025) study sequential joint pricing–inventory policies. The closest online work is Han et al. (2026a), which represents demand using known basis functions with unknown coefficients and also analyzes approximation error. Its algorithm uses an initial high-inventory exploration phase to obtain demand observations before committing to an estimated policy. We study learning guarantees over nonparametric smoothness classes in context and price. Our policy requires visibility of a low calibration quantile rather than full demand, and continues to update price estimates after local inventory profiling. Our analysis focuses on within-period stocking decisions, whereas Han et al. (2026a) also account for inventory carryover. For this setting, we establish matching upper and lower regret bounds that characterize the dependence on context dimension, contextual smoothness, and price-value regularity.

3 Problem Setup

Decisions and feedback. Let $\mathcal{X} = [0,1]^d$ be the context space and $\mathcal{P} = [p_{\min}, p_{\max}]$, with $0 < p_{\min} < p_{\max} < \infty$, the price range. The seller chooses actions from $\mathcal{A} = \mathcal{P} \times [y_{\min}, y_{\max}]$, where $0 \le y_{\min} < y_{\max} < \infty$. The action ranges and per-unit lost-sales and holding costs $b, h_c > 0$ are known.

In each period $t = 1, \dots, T$, the seller observes x_t and chooses a price p_t and inventory y_t. A policy π maps the current context and observed history $\mathcal{H}_{t-1} = \{(x_s, p_s, y_s, O_s) : s < t\}$ to a possibly randomized action $(p_t, y_t) \in \mathcal{A}$. Inventory is replenished for each period, with no carryover. Demand then realizes as

$$D_t = \lambda(x_t, p_t) + \epsilon_t,$$

where the demand location λ and residual distribution are unknown. The seller observes sales $O_t = \min\{D_t, y_t\}$. Sales reveal demand when $O_t < y_t$. A stockout, identified by $O_t = y_t$, reveals only that $D_t \ge y_t$.

Profit and learning objective. The one-period profit is

$$r((p,y), D) = p \min\{D, y\} - b(D-y)^+ - h_c(y-D)^+.$$

Lost-sales costs depend on unmet demand, so realized profit is partially unobserved when a stockout occurs. For context x, write the expected profit as

$$Q_x(p,y) = \mathbb{E}[r((p,y), D) \mid x, p].$$

We compare the policy with a seller who knows the conditional demand distribution and maximizes expected profit for each observed context. This benchmark chooses its action before demand realizes. Since inventory is replenished between periods, its optimization separates across periods. Optimizing inventory at a given price and then choosing the price gives

$$G_x(p) = \max_{y \in [y_{\min}, y_{\max}]} Q_x(p,y), \qquad V_x = \max_{p \in \mathcal{P}} G_x(p).$$

Regret is the expected cumulative profit loss relative to this benchmark:

$$\mathsf{Reg}_T(\pi) = \mathbb{E}_\pi \left[\sum_{t=1}^{T} \{V_{x_t} - Q_{x_t}(p_t, y_t)\} \right].$$

The expectation is over contexts, demand shocks, and policy randomization. The objective is to minimize this loss using the censored sales generated by the policy's own decisions.

Demand structure and regularity. We first specify how contexts and demand shocks arrive. We then impose smoothness for sharing information across contexts and prices, followed by conditions on the quantiles used for inventory learning and calibration.

Assumption 1 (Demand process)**.** ***(i) Arrivals.*** *The pairs (x_t, ϵ_t) are i.i.d. across periods and independent of policy randomization. Contexts follow an arbitrary distribution on $\mathcal{X}$.*

(ii) Conditional shocks. *Given $x_t = x$, the shock ϵ_t has distribution F_x, with mean zero, a density f_x, and common compact support $[\underline{z}, \overline{z}]$. The location surface λ is uniformly bounded, and demand is nonnegative almost surely.*

Thus $\lambda(x,p) = \mathbb{E}[D \mid x,p]$. At a fixed context, prices shift the demand location while the residual distribution stays fixed. Sharing observations across contexts also requires control of changes in this distribution.

Assumption 2 (Context and price smoothness)**.** ***(i) Context variation.*** *For some $0 < \alpha \leq 1$ and finite L_x, L_F, uniformly in $x, x' \in \mathcal{X}$ and $p \in \mathcal{P}$,*

$$|\lambda(x,p) - \lambda(x',p)| \leq L_x \|x - x'\|^\alpha,$$
$$\sup_{u\in\mathbb{R}} |F_x(u) - F_{x'}(u)| \leq L_F \|x - x'\|^\alpha.$$

(ii) Price variation. *For a finite L_p, $|\lambda(x,p) - \lambda(x,q)| \leq L_p|p-q|$ uniformly in $x \in \mathcal{X}$ and $p, q \in \mathcal{P}$.*

Inventory learning uses the demand quantile associated with the critical ratio, while price calibration uses a lower quantile. To describe their regularity, define $F_x^{-1}(u) = \inf\{z : F_x(z) \geq u\}$ for $u \in (0,1)$, and write

$$z_u(x) = F_x^{-1}(u), \qquad q_u(x,p) = \lambda(x,p) + z_u(x), \qquad \phi(p) = \frac{p+b}{p+b+h_c},$$

and abbreviate $z_p(x) = z_{\phi(p)}(x)$. Fix a calibration level $\rho \in (0, \inf_{p\in\mathcal{P}} \phi(p))$.

Assumption 3 (Quantile regularity and inventory feasibility)**.** ***(i) Density bounds.*** *There are $\kappa, \kappa', r_q > 0$ such that $0 \leq f_x(u) \leq \kappa'$ on the common support, uniformly in x. Moreover, for every x and p,*

$$f_x(u) \geq \kappa \quad \textit{whenever} \quad |u - z_\rho(x)| \leq r_q \quad \textit{or} \quad |u - z_p(x)| \leq r_q.$$

(ii) Inventory margins. *There is $r_y \in (0, r_q]$ such that, uniformly in x, p,*

$$y_{\min} + r_y \leq q_\rho(x,p), \qquad q_{\phi(p)}(x,p) \leq y_{\max} - r_y.$$

These conditions yield the unique optimal inventory $y_x^*(p) = q_{\phi(p)}(x,p)$ and a uniform gap $y_x^*(p) - q_\rho(x,p) \geq m_q > 0$ (Lemmas 2 and 3). This gap allows inventory estimates close to $y_x^*(p)$ to expose the calibration quantile in sales. The price smoothness in Assumption 2 also implies that G_x is uniformly Lipschitz (Lemma 4). The faster-rate specialization in Section 5 imposes additional smoothness on G_x.

4 Learning and Deciding with a Surrogate Reward

Sales censoring makes unfulfilled demand unobservable, yet the occurrence of a stockout remains observable. Furthermore, censoring only truncates the upper tail of the demand distribution without distorting the shape of its observable, uncensored portion. Motivated by these observations, we propose a two-tier surrogate evaluation framework that preserves expected profit differences for inventory and price comparisons within a fixed context.

Since exact context repetitions are rare in a continuous context space, a data-driven approach must draw on observations from neighboring contexts. Doing so introduces a trade-off between local accuracy and data sufficiency, alongside the exploration-exploitation trade-off inherent in sequential learning. A kernel-based UCB framework naturally addresses both challenges. Moreover, the structural relationship between price and inventory allows inventory information to be shared across nearby prices and contexts, reducing exploration costs while controlling the resulting expected profit loss.

Building on these insights, we introduce the Surrogate-Profit UCB (SP-UCB) algorithm.

The policy is supplied with T, the context smoothness exponent α, the calibration level ρ, and the choice of the Lipschitz or smooth-price regime. We select a kernel and tuning constants using known numerical bounds for the model class. The required bounds and parameter choices are specified in Appendix C.1.

4.1 Constructing a Two-Tier Surrogate Profit from Observable Data

For a fixed context, we solve the censoring problem in two steps. First, an observable sales-based surrogate differs from the expected profit by an offset. Since this offset ignores inventory, the surrogate perfectly captures the profit gaps between different inventory choices at a fixed price. Second, although this offset changes across prices, an observable low-demand quantile shifts with price by the exact same amount as mean demand. We obtain a price-independent offset by subtracting this quantile multiplied by the per-unit lost-sales cost (b). Ultimately, these two steps create a two-tier surrogate that evaluates inventory levels first and prices second.

We decouple the true profit into an observable component and an unobservable remainder. For price p, inventory y, and observed sales $O = \min\{D, y\}$, where D is demand, define surrogate score $Y(p, y, O)$ from observable information

$$Y(p, y, O) := (p + b)O - h_c(y - O),$$

where b and h_c are the per-unit lost-sales and holding costs. The realized profit can be reformulated as $r((p, y), D) = Y(p, y, O) - bD$. Let $\overline{Q}_x(p, y)$ denote the expected surrogate score at context x. Its expectation is

$$\mathbb{E}[Y(p, y, O) \mid x, p, y] = \overline{Q}_x(p, y) = Q_x(p, y) + b\lambda(x, p),$$

where $Q_x(p, y)$ is the expected profit and $\lambda(x, p) = \mathbb{E}[D \mid x, p]$ by the demand model and the zero-mean noise condition in Assumption 1. Thus, finding the price–inventory pair with the highest expected profit can be approached by computing Y from observable data, obtaining information about the mean-demand offset $b\lambda(x, p)$, and maximizing the corrected value.

For a given context x, each price p admits an optimal inventory level $y_x^*(p)$ that maximizes the expected surrogate score. The mean-demand offset $b\lambda(x, p)$ is independent of inventory decisions. Therefore, although the full profit level remains unknown, expected profit differences between inventory decisions at a fixed context and price, including the loss from deviating from the optimal inventory, can already be computed through the surrogate, as detailed in Proposition 5.

Let $S_x(y \mid p) := \mathbb{P}(D \geq y \mid x, p)$ denote the stockout probability and $\tau(p) := h_c/(p + b + h_c)$ the critical stockout threshold (Arrow et al., 1951).

Then

$$\partial_y \mathbb{E}[Y(p, y, O) \mid x, p] = (p + b + h_c)\big(S_x(y \mid p) - \tau(p)\big).$$

Since $S_x(y \mid p)$ is nonincreasing in y, so is this derivative. When the stockout probability exceeds the critical threshold, increasing inventory improves expected profit; when it is below the threshold, decreasing inventory improves expected profit. Moreover, controlling the gap $|S_x(y \mid p) - \tau(p)|$ controls the inventory error locally, as formalized below.

Lemma 1 (Inventory optimality). *The optimal inventory is $y_x^*(p) = \lambda(x, p) + z_{\phi(p)}(x)$. It uniquely solves $S_x(y \mid p) = \tau(p)$ and maximizes $\mathbb{E}[Y(p, y, O) \mid x, p]$. There exist constants $c_y, C_y > 0$ such that, whenever $|y - y_x^*(p)| \leq r_y$,*

$$c_y|y - y_x^*(p)| \leq |S_x(y \mid p) - \tau(p)| \leq C_y|y - y_x^*(p)|,$$

and

$$G_x(p) - Q_x(p, y) \leq C_y|y - y_x^*(p)|^2.$$

The analysis above characterizes the optimal inventory for a fixed context and price. However, learning in a continuous context space cannot generally rely on repeated observations at exactly the same context, motivating the use of data from nearby contexts.

We partition the context space into a context grid and pool observations within each cell to learn its inventory decision.

Definition 1 (Context grid). *A context grid at resolution $\delta_x > 0$ is a finite measurable partition $\mathcal{C}$ of $\mathcal{X}$ satisfying*

$$\operatorname{diam}(C) \leq \delta_x \quad \textit{for every } C \in \mathcal{C}, \qquad |\mathcal{C}| \leq C_{\text{cell}}\delta_x^{-d},$$

where C_{cell} is independent of δ_x.

A sufficiently fine context grid allows a single inventory decision to approximate the optimal inventory throughout each cell (Lemma 7). We therefore propose a bisection procedure to learn the inventory target within each context cell. At each selected price, the procedure compares the pooled empirical stockout probability with the critical threshold to determine the direction of inventory adjustment. The full procedure is given in Algorithm 2.

Proposition 1 (Inventory learning)**.** *Fix a price p, a target value accuracy $\varepsilon > 0$, and a context grid with $\delta_x^\alpha \le \sqrt{\varepsilon}$. For sufficiently small ε, with probability at least $1-\delta$, the estimates returned by Algorithm 2 satisfy, simultaneously for every completed cell C,*

$$\sup_{x \in C} |\widehat{y}_C(p) - y_x^*(p)| \le C_{\mathrm{inv}} \sqrt{\varepsilon},$$

where $C_{\mathrm{inv}} > 0$ depends only on the model constants and fixed algorithm parameters, not on T, ε, or δ_x.

The procedure uses at most $\widetilde{O}(\varepsilon^{-1})$ rounds per cell. The cumulative regret incurred during inventory search in each completed cell is at most $\widetilde{O}\big(\varepsilon^{-1}\big)$.

Inventory profiling also enables price comparison by making the low quantile $q_\rho(x,p)$ observable below the certified inventory $\widehat{y}_C(p)$ (Lemma 6). Although this inventory already supports stocking decisions at a fixed price, comparing prices still requires correcting the price-dependent offset $b\lambda(x,p)$ in the surrogate score. One intuitive approach is to estimate $\lambda(x,p)$ empirically by setting inventory relatively high to reduce lost sales. However, this approach incurs additional inventory costs, and the feasible inventory range may not be sufficient to reveal all demand. Instead, the visible low quantile shifts with price by exactly the same amount as mean demand at a fixed context. It therefore provides the required correction without recovering the unobserved demand tail.

Recall that the expected surrogate score takes the form $\mathbb{E}[Y(p,y,O) \mid x,p] = Q_x(p,y) + b\lambda(x,p)$. The demand model implies $q_\rho(x,p) = \lambda(x,p) + z_\rho(x)$, where $z_\rho(x)$ is the ρ-quantile of the price-independent residual noise. Subtracting $bq_\rho(x,p)$ from the expected surrogate isolates the price-dependent terms:

$$\mathbb{E}[Y(p,y,O) \mid x,p] - bq_\rho(x,p) = Q_x(p,y) - bz_\rho(x).$$

Crucially, the remaining offset $bz_\rho(x)$ is invariant to price and thus preserves the true profit landscape for price comparison. Building on this price-invariance, we evaluate prices using the optimized surrogate value $\overline{G}_x(p) := \max_y \mathbb{E}[Y(p,y,O) \mid x,p]$. We define the corrected surrogate profit function as:

$$H_x(p) := \overline{G}_x(p) - bq_\rho(x,p).$$

It follows that $H_x(p) = G_x(p) - bz_\rho(x)$. Because $H_x(p)$ and the true expected profit $G_x(p)$ differ only by a context-dependent constant, they share the exact same set of optimal prices. Maximizing $H_x(p)$ thus recovers the optimal price, which can then be paired with its certified inventory level to yield a complete joint decision.

4.2 Maximizing the Surrogate Profit with Contextual Information

Although contextual smoothness provides limited guarantees for transferring information across nearby contexts, effective use of contextual data requires a trade-off between the amount of data available in the neighborhood and the accuracy of the resulting information, which is further compounded by the exploration-exploitation trade-off inherent in sequential learning.

To navigate these challenges, we propose a kernel-based Upper Confidence Bound (UCB) framework. To compare candidate prices using UCB, we restrict our search for the optimal price to a finite candidate price set Γ, specified in Definition 2 in the appendix. An intuitive approach is to run Inventory Search (Algorithm 2) separately for every candidate price. However, inventory learning need only be accurate enough to support price comparison. Near the optimal inventory, small inventory errors have only a second-order effect on expected profit. Since the optimal inventory also varies smoothly with price, an inventory learned at one price can serve nearby prices while maintaining the desired profit accuracy.

Exploiting this structure, we perform inventory search on a coarser price grid Γ_y, and let each candidate price $p \in \Gamma$ use the inventory learned at its nearest grid point in Γ_y.

Proposition 2. *For Lipschitz price values and sufficiently small ε, applying Algorithm 2 on $\Gamma = \Gamma(\varepsilon)$ or $\Gamma_y = \Gamma(\sqrt{\varepsilon})$ yields*

$$\sup_{x \in C,\ p \in \Gamma} \left\{ G_x(p) - Q_x(p, \widehat{y}_C(p)) \right\} \le C_{\text{share}} \varepsilon$$

for every completed cell C with probability at least $1 - \delta$. $C_{\text{share}} > 0$ is independent of T and ε. The respective per-cell inventory-search regret bounds are $\widetilde{O}(\varepsilon^{-2})$ for Γ, $\widetilde{O}(\varepsilon^{-3/2})$ for Γ_y.

With the shared inventory decisions in place, we estimate the surrogate profit at the current context by leveraging nearby observations. A kernel restricts this pooling to a neighborhood whose bandwidth controls the localization bias. Let $h > 0$ denote the bandwidth and let $K : \mathbb{R}^d \to [0, 1]$ be a compactly supported Lipschitz kernel, with $K(u) = 0$ for $\|u\| > 1$ and $K(u) \ge k_0 > 0$ for $\|u\| \le 1/2$.

For each $p \in \Gamma$, let $\mathcal{T}_t(p)$ contain the rounds before t in which price p was selected after Algorithm 2 completed the inventory search for all prices in Γ_y within the corresponding context cell. Let x_s denote the context observed in round s, and let x_t denote the current context. For each $s \in \mathcal{T}_t(p)$, the normalized weight is

$$\frac{K((x_s - x_t)/h)}{W_t(x_t, p)}, \qquad W_t(x_t, p) := \sum_{s \in \mathcal{T}_t(p)} K((x_s - x_t)/h),$$

which is defined for $W_t(x_t, p) > 0$.

Under this weighting scheme, Let $\widehat{\overline{G}}_t(x_t, p)$ be the weighted average of observed surrogate scores, and $\widehat{q}_{\rho,t}(x_t, p)$ be the ρ-quantile of the empirical sales distribution formed with the same weights. These two estimates approximate the surrogate value and demand quantile at the current context. With sufficient amount of neighboring observations, their difference,

$$\widehat{H}_t(x_t, p) := \widehat{\overline{G}}_t(x_t, p) - b\widehat{q}_{\rho,t}(x_t, p),$$

approximates the learning target $H_x(p)$.

The estimate $\widehat{H}_t(x_t, p)$ is subject to three sources of error: the profit loss from using learned and shared inventories, the localization bias from pooling nearby contexts, and the sampling uncertainty in the weighted observations. We account for these errors through a confidence width and select prices optimistically. This allows a sparsely sampled price to remain competitive even when its current estimated value is low. For contextual Hölder exponent α, choose a bandwidth $h > 0$ satisfying $h^\alpha = \varepsilon$. For a sufficiently large fixed constant $C > 0$, define the confidence width as

$$B_t(x_t, p) := C \left[\varepsilon + h^\alpha + \sqrt{\frac{\log(T|\Gamma|/h)}{\max\{1, W_t(x_t, p)\}}} \right].$$

The upper confidence index is

$$U_t(x_t, p) := \widehat{H}_t(x_t, p) + B_t(x_t, p),$$

with $U_t(x_t, p) = +\infty$ for $W_t(x_t, p) = 0$.

Once the inventory search is complete for Γ_y within the current context cell, the algorithm selects $p_t \in \arg\max_{p \in \Gamma} U_t(x_t, p)$ and uses the inventory associated with the nearest grid point to p_t in Γ_y. The resulting sales observation updates the surrogate profit estimate and confidence width for that price at nearby contexts, thereby refining the upper confidence indices used in subsequent rounds. The complete SP-UCB algorithm is presented in Algorithm 1.

Detailed error analysis, algorithmic procedures, and parameter choices are provided in the appendix Appendices B, C.1, E and F.

5 Main Results

Let $\mathfrak{I}_{\text{Lip}}$ denote the instance class satisfying Assumptions 1 to 3, with fixed numerical bounds on the model constants. Its price values are uniformly Lipschitz by Lemma 4. A smoother price-value curve yields a faster rate under the following additional condition.

Assumption 4 (Smooth-price specialization). *For every x, G_x is twice differentiable in price, with $\sup_{x,p} |\partial_{pp} G_x(p)| \le L_{G,2}$ for a fixed finite $L_{G,2}$.*

Let $\mathfrak{I}_{\mathrm{sm}}$ be the subclass satisfying Assumption 4. The formal upper and lower statements, including the fixed numerical envelopes and tuning constants, are given in Appendix C. We use the bandwidth h, target value accuracy $\varepsilon = h^{\alpha}$, and candidate price set Γ from Section 4. The fixed multiplier $c_\eta > 0$ sets the stockout-probability tolerance $\eta = c_\eta \sqrt{\varepsilon}$ in the inventory search, as specified in Appendix C.1.

Theorem 1 (Upper bounds for contextual censored sales). *For fixed $b, h_c > 0$ and any fixed $c_\eta > 0$, with certified tuning as in Appendix C.1,* SP-*UCB satisfies, up to logarithmic factors,*

$$\sup_{\mathcal{I} \in \mathfrak{I}_{\mathrm{Lip}}} \mathsf{Reg}_T(\mathsf{SP\text{-}UCB}; \mathcal{I}) \le \widetilde{O}\Big(T^{\frac{d+2\alpha}{d+3\alpha}}\Big), \qquad \sup_{\mathcal{I} \in \mathfrak{I}_{\mathrm{sm}}} \mathsf{Reg}_T(\mathsf{SP\text{-}UCB}; \mathcal{I}) \le \widetilde{O}\Big(T^{\frac{2d+3\alpha}{2d+5\alpha}}\Big).$$

Remark 1. *The multiplier c_η is fixed independently of T, and the hidden constants may depend on it. The upper bound follows the learning steps in Section 4. The algorithm first searches for inventories within context cells. Once this search is complete in a cell, the returned inventories make the low quantile observable, allowing the algorithm to estimate $H_x(p)$ and compare candidate prices using $U_t(x_t, p)$. Since $H_x(p)$ preserves price gaps, errors in this comparison translate into expected profit loss. Suppressing logarithms and constants, the regret bound has the form*

$$h^{-d/2}\varepsilon^{-3/2} \;+\; T\varepsilon \;+\; \sqrt{T|\Gamma|h^{-d}}.$$

The first term bounds the cost of inventory search over the context cells and inventory-search prices specified in Appendix B. The second collects inventory approximation, context localization, and price discretization errors, each of order ε per subsequent round. The third bounds accumulated statistical uncertainty in the kernel estimates. By Definition 2, $|\Gamma| \asymp \varepsilon^{-1}$ for Lipschitz price values and $|\Gamma| \asymp \varepsilon^{-1/2}$ for twice-smooth price values.

Theorem 2 (Lower bounds for contextual censored sales). *For fixed $b, h_c > 0$, every policy incurs, up to logarithmic factors,*

$$\inf_{\pi} \sup_{\mathcal{I} \in \mathfrak{I}_{\mathrm{Lip}}} \mathsf{Reg}_T(\pi; \mathcal{I}) \ge \widetilde{\Omega}\Big(T^{\frac{d+2\alpha}{d+3\alpha}}\Big), \qquad \inf_{\pi} \sup_{\mathcal{I} \in \mathfrak{I}_{\mathrm{sm}}} \mathsf{Reg}_T(\pi; \mathcal{I}) \ge \widetilde{\Omega}\Big(T^{\frac{2d+3\alpha}{2d+5\alpha}}\Big).$$

Remark 2. *The lower bounds match the upper-bound exponents over the stated instance classes. They show that no policy using the same censored-sales feedback can achieve a smaller worst-case polynomial order of regret. The constructions and information bounds are given in Appendix C. This comparison establishes the optimality of the overall rates; it does not require every policy to use the same inventory search or incur each term of the upper-bound decomposition.*

Interpreting the exponents. A smaller bandwidth h reduces localization bias but leaves fewer nearby observations for each price. A smaller target value error ε also requires more accurate inventory decisions and a finer candidate price set. With $\varepsilon = h^{\alpha}$, balancing $T\varepsilon$ and $\sqrt{T|\Gamma|h^{-d}}$ gives

$$h \asymp T^{-1/(d+3\alpha)} \quad \text{for Lipschitz price values,} \qquad h \asymp T^{-2/(2d+5\alpha)} \quad \text{for twice-smooth price values,}$$

up to logarithmic factors. At these choices, the inventory-search term is of lower order than the resulting regret bound. Larger d worsens the exponent because observations are spread across more context neighborhoods, while larger α allows more sharing with less bias. The twice-smooth specialization improves the rate because a price mesh of order $\sqrt{\varepsilon}$ still incurs only $O(\varepsilon)$ optimized-value loss, reducing the number of prices that must be compared.

6 Experiments

We visualize the finite-horizon cumulative regret of SP-UCB across different context dimensions and smoothness levels. The main experiment uses a controlled demand model in which the demand location varies smoothly with context and price, while the residual noise follows a uniform distribution. We run five independent replications for each of six configurations, covering one- and two-dimensional contexts and smoothness levels from $\alpha = 0.5$ to the Lipschitz case $\alpha = 1$, at horizons $T \in \{4000, 8000, 16000, 32000, 64000\}$. The resulting curves are finite-sample diagnostics of the scaling behavior. Full hyperparameters and implementation details are in Appendix I.

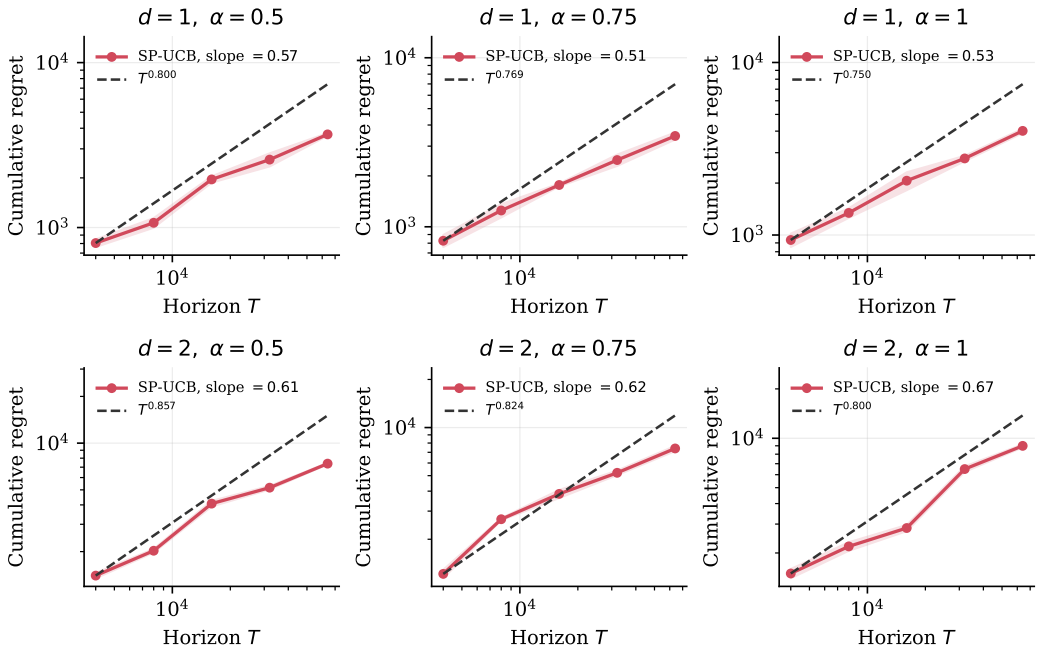


Figure 1: Cumulative-regret scaling of the paper-aligned SP-UCB policy for different context dimensions d and Hölder exponents α. Solid curves are means over five exogenous trajectories, and dashed curves give representative $T^{\beta(d,\alpha)}$ guides normalized at the first plotted horizon.

The fitted log–log slopes are $(0.57, 0.51, 0.53)$ for $d = 1$ and $(0.61, 0.62, 0.67)$ for $d = 2$, in increasing order of smoothness. These fits summarize the five tested horizons and do not estimate a worst-case asymptotic exponent. When $d = 2$, profiling occupies a larger fraction of the rounds and the local histories are sparser. The resulting curves should therefore be read as finite-sample diagnostics of the profiling and UCB interaction, rather than as a verification of the minimax rate.

Additional experiments Appendix I compares SP-UCB with neighboring context-pooled methods on a two-dimensional instance with nonlinear context interactions in demand and a shared uniform residual distribution. It also reports the twice-smooth price specialization and a smoothness-mismatch study that varies the smoothness parameter supplied to SP-UCB while holding the data-generating environment fixed.

7 Conclusion

We studied a seller who must make two decisions at once: how much to charge and how much to stock. The seller sees the context and the resulting sales, but not the demand that was turned away. We showed that, under a nonparametric contextual location model, the visible part of censored sales can still support joint pricing and stocking. SP-UCB combines inventory profiling, local contextual pooling, and calibrated price learning, and achieves matching minimax rates for the stated Lipschitz and twice-smooth price classes.

The result remains a per-period model without inventory carryover or lead times, and the shared-noise location structure does not cover arbitrary price-dependent noise distributions. Extensions to dynamic inventory states, weaker feedback, or a more general conditional demand law are natural directions for future work.

REFERENCES

Kenneth J. Arrow, Theodore Harris, and Jacob Marschak. Optimal inventory policy. *Econometrica*, 19(3):250–272, 1951. doi: 10.2307/1906813.

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. doi: 10.1287/mnsc.2020.3680.

Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2019. doi: 10.1287/opre.2018.1757.

Omar Besbes and Alp Muharremoglu. On implications of demand censoring in the newsvendor problem. *Management Science*, 59(6):1407–1424, 2013. doi: 10.1287/mnsc.1120.1654.

Jinzhi Bu, David Simchi-Levi, and Li Wang. Offline pricing and demand learning with censored data. *Management Science*, 69(2):885–903, 2023. doi: 10.1287/mnsc.2022.4382.

Felipe Caro and Jérémie Gallien. Clearance pricing optimization for a fast-fashion retailer. *Operations Research*, 60(6):1404–1422, 2012. doi: 10.1287/opre.1120.1102.

Boxiao Chen, Xiuli Chao, and Cong Shi. Nonparametric learning algorithms for joint pricing and inventory control with lost sales and censored demand. *Mathematics of Operations Research*, 46 (2):726–756, 2021. doi: 10.1287/moor.2020.1084.

Boxiao Chen, Yining Wang, and Yuan Zhou. Optimal policies for dynamic pricing and inventory control with nonparametric censored demands. *Management Science*, 70(5):3362–3380, 2024. doi: 10.1287/mnsc.2023.4859.

Ningyuan Chen and Guillermo Gallego. Nonparametric pricing analytics with customer covariates. *Operations Research*, 69(3):974–984, 2021. doi: 10.1287/opre.2020.2016.

Ningyuan Chen and Guillermo Gallego. A primal–dual learning algorithm for personalized dynamic pricing with an inventory constraint. *Mathematics of Operations Research*, 47(4):2585–2613, 2022. doi: 10.1287/moor.2021.1220.

Maxime C. Cohen, Ilan Lobel, and Renato Paes Leme. Feature-based dynamic pricing. *Management Science*, 66(11):4921–4943, 2020. doi: 10.1287/mnsc.2019.3485.

Christopher T. Conlon and Julie Holland Mortimer. Demand estimation under incomplete product availability. *American Economic Journal: Microeconomics*, 5(4):1–30, 2013. doi: 10.1257/mic.5.4.1.

Jingying Ding, Woonghee Tim Huh, and Ying Rong. Feature-based inventory control with censored demand. *Manufacturing & Service Operations Management*, 26(3):1157–1172, 2024. doi: 10.1287/msom.2021.0135.

Kris Johnson Ferreira, Bin Hong Alex Lee, and David Simchi-Levi. Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*, 18(1):69–88, 2016. doi: 10.1287/msom.2015.0561.

Dylan J. Foster, Yanjun Han, Jian Qian, and Alexander Rakhlin. Online estimation via offline estimation: An information-theoretic framework. In *Advances in Neural Information Processing Systems*, volume 37, pp. 42840–42898, 2024. doi: 10.52202/079017-1357.

Korel Gundem and Zhengling Qi. Offline dynamic inventory and pricing strategy: Addressing censored and dependent demand. *arXiv preprint arXiv:2504.09831*, 2025.

Zean Han, Zezhen Ding, and Jiheng Zhang. Learning to price and stock under contextual and censored demand. In Diego Calvanese (ed.), *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, pp. 4339–4347. International Joint Conferences on Artificial Intelligence Organization, 8 2026a. doi: 10.24963/ijcai.2026/483. URL https://doi.org/10.24963/ijcai.2026/483. Main Track.

Zean Han, Ruihan Lin, Zezhen Ding, and Jiheng Zhang. Direction-aware offline-to-online learning in linear contextual bandits. *arXiv preprint arXiv:2604.24016*, 2026b. doi: 10.48550/arXiv.2604.24016. URL https://arxiv.org/abs/2604.24016.

Woonghee Tim Huh and Paat Rusmevichientong. A nonparametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research*, 34(1):103–123, 2009. doi: 10.1287/moor.1080.0355.

Woonghee Tim Huh, Retsef Levi, Paat Rusmevichientong, and James B. Orlin. Adaptive data-driven inventory control with censored demand based on Kaplan–Meier estimator. *Operations Research*, 59(4):929–941, 2011. doi: 10.1287/opre.1100.0906.

Adel Javanmard and Hamid Nazerzadeh. Dynamic pricing in high-dimensions. *Journal of Machine Learning Research*, 20(9):1–49, 2019. URL http://jmlr.org/papers/v20/17-357.html.

Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. *Journal of Machine Learning Research*, 21(137):1–45, 2020.

Ruihan Lin, Yunrui Ling, and Jiheng Zhang. Dynamic pricing under revenue-based financing: Localized and uniformly bounded inefficiency. SSRN working paper, 2026. URL https://ssrn.com/abstract=6237799.

Andrés Musalem, Marcelo Olivares, Eric T. Bradlow, Christian Terwiesch, and Daniel Corsten. Structural estimation of the effect of out-of-stocks. *Management Science*, 56(7):1180–1197, 2010. doi: 10.1287/mnsc.1100.1170.

Hanzhang Qin, David Simchi-Levi, and Li Wang. Data-driven approximation schemes for joint pricing and inventory control models. *Management Science*, 68(9):6591–6609, 2022. doi: 10.1287/mnsc.2021.4212.

Aleksandrs Slivkins. Contextual bandits with similarity information. *Journal of Machine Learning Research*, 15(73):2533–2568, 2014.

Jingwen Tang, Zhengling Qi, Ethan Fang, and Cong Shi. Offline feature-based pricing under censored demand: A causal inference approach. *Manufacturing & Service Operations Management*, 27(2):535–553, 2025. doi: 10.1287/msom.2024.1061.

Yining Wang, Boxiao Chen, and David Simchi-Levi. Multimodal dynamic pricing. *Management Science*, 67(10):6136–6152, 2021. doi: 10.1287/mnsc.2020.3819.

A Proof Roadmap and Common Notation

The proofs proceed through identification, estimation, regret accounting, and hardness. With only censored sales available to the learner, the calibrated value provides an estimable target:

$$H_x(p) := \overline{G}_x(p) - bq_\rho(x,p) = G_x(p) - bz_\rho(x).$$

The first equality identifies a target that can be estimated from sales, and the second shows that its maximizers are exactly those of G_x. The upper-bound proof proceeds in three steps. We first bound the inventory error and establish quantile visibility. We then control the calibrated estimation error and sum the resulting confidence widths. The supporting results are organized as follows:

1. Proposition 5 and Lemma 5 establish that a censored observation supplies a bounded proxy sample and a visible quantile correction for $H_x(p)$.
2. Lemma 6 profiles only coarse context cells and $\sqrt{\varepsilon}$-spaced price anchors, then transfers the nearest anchor estimate to each decision price. The resulting inventory is within $O(\sqrt{\varepsilon})$ of the critical inventory. Its value loss is $O(\varepsilon)$, and its positive separation above $q_\rho(x,p)$ makes the calibration quantile uncensored.
3. Lemmas 8 and 9 give a simultaneous confidence interval for $H_x(p)$ under adaptive price choices. The radius contains a stochastic term $W_t(x,p)^{-1/2}$ and the two bias terms ε and h^α.
4. Lemma 10 sums the stochastic radii. Price discretization and profiling supply the remaining deterministic regret terms, producing the three-term decomposition used in Theorem 3 and Corollary 1.

For the lower bound, we work with the admissible subfamily $F_x \equiv F_0$ and reveal the common noise law to the learner. A value bump $g(x,p)$ is implemented by the location shift $g(x,p)/p$. Lemma 12 verifies that these shifts remain inside the model class. The square-root density of the fixed noise law has finite translation energy, so a location shift of size $O(\Delta)$ contributes only $O(\Delta^2)$ squared Hellinger information. The complete proof in Appendix H averages this information over disjoint price supports inside every context cell and then applies a product prior over cell labels. This is the direct-sum step responsible for the factor h^{-d}.

The common compact support of ϵ, boundedness of λ, and compact action sets imply constants $Y_{\max}, R_{\max} < \infty$ such that $|Y_t| \le Y_{\max}$ and one-round regret is at most $R_{\max}$ on every instance considered below. Constants denoted by C, c may change from line to line and depend only on the fixed model constants, never on T, h, ε, or the packing labels.

B Algorithm Listings

Definition 2 (Candidate price set)**.** *For $\Delta > 0$, let $\Gamma(\Delta)$ denote the uniform price grid on $\mathcal{P}$ with mesh at most Δ, including both endpoints. For a target value accuracy $\varepsilon > 0$, define*

$$\Gamma := \Gamma_{\mathrm{dec}} = \Gamma(\Delta_{\mathrm{dec}}), \qquad \Delta_{\mathrm{dec}} := \begin{cases} \varepsilon, & \textit{under Lipschitz price-value regularity}, \\ \sqrt{\varepsilon}, & \textit{under the smooth-price specialization}. \end{cases}$$

The two cases use the Lipschitz property in Lemma 4 and the additional smoothness in Assumption 4. Both yield an $O(\varepsilon)$ approximation error in optimized expected profit by Proposition 4.

For a context bandwidth h and target value accuracy $\varepsilon = h^\alpha$, we therefore profile on the coarser resolutions

$$\delta_x = \varepsilon^{1/(2\alpha)} = \sqrt{h}, \qquad \Gamma_y = \Gamma(\sqrt{\varepsilon}).$$

Partition the context space into cells $\{C_z^{\mathrm{prof}} : z \in \mathcal{Z}_{\mathrm{prof}}\}$ of diameter at most δ_x, with representative z and $|\mathcal{Z}_{\mathrm{prof}}| \le C\delta_x^{-d}$. Each cell–anchor pair (C_z^{prof}, a), $a \in \Gamma_y$, has its own bisection state. A decision price p uses the inventory returned for its nearest anchor $\pi(p)$.

We separate the profiling precision from these geometric resolutions by setting $\eta = c_\eta\sqrt{\varepsilon}$, where $c_\eta > 0$ is a fixed, horizon-independent input. The profiler compares empirical stockout rates with $\tau(a) \pm 2\eta$ and uses batches proportional to η^{-2}, up to logarithmic factors. Decreasing c_η narrows the

stopping band but requires more samples. It leaves the context partition and anchor grid unchanged. The choice $c_\eta = 1$ gives the unit-multiplier setting.

Context arrivals determine when these local estimates can be formed. Because Assumption 1 imposes no lower bound on cell probabilities, requiring all cells to finish profiling would delay price learning in cells that already have enough observations. The policy updates the unfinished profiler of the arriving cell and begins UCB immediately upon that cell's completion. The persistent states and one-round update are given in Algorithm 2.

Under the certified resolution $h \le h_{\mathrm{vis}}$, Lemma 6 guarantees, with high probability, that every completed cell returns inventories with $O(\sqrt{\varepsilon})$ error, $O(\varepsilon)$ value loss, and a visible ρ-quantile. The total number of profiling rounds is deterministically bounded by $\widetilde{O}(h^{-d/2}\varepsilon^{-3/2})$. These guarantees allow each completed cell to use subsequent sales for both terms of $H_x(p)$.

Algorithm 1 Surrogate-Profit UCB (SP-UCB)

Require: horizon T, context smoothness α, visible quantile level ρ, price mode $s \in \{\mathrm{Lip}, \mathrm{smooth}\}$, fixed tuning constants $C_{\mathrm{tune}}, h_{\mathrm{vis}}, c_\eta > 0$

Fix $a_{\mathrm{fb}} = (p_{\min}, y_{\min})$ and set $\delta = T^{-2}$
if $s = \mathrm{Lip}$ **then**
    Set $h = (\log(eT)/T)^{1/(d+3\alpha)}$, $\varepsilon = h^\alpha$, and $\Delta_{\mathrm{dec}} = \varepsilon$
else
    Set $h = (\log(eT)/T)^{2/(2d+5\alpha)}$, $\varepsilon = h^\alpha$, and $\Delta_{\mathrm{dec}} = \sqrt{\varepsilon}$
Set $\Gamma_{\mathrm{dec}} = \Gamma(\Delta_{\mathrm{dec}})$, $\Gamma_y = \Gamma(\sqrt{\varepsilon})$, and $\delta_x = \varepsilon^{1/(2\alpha)}$. Construct $\mathcal{Z}_{\mathrm{prof}}$ with cell diameter at most δ_x, and initialize the persistent anchor-profiler states in Algorithm 2 with $\eta = c_\eta\sqrt{\varepsilon}$
if $h > h_{\mathrm{vis}}$ **then**
    Play a_{fb} on all T rounds and **stop**
for round $t = 1, \dots, T$ **do**
    Observe context x_t and let $z = z_{\mathrm{prof}}(x_t)$
    if cell z is not profiled **then**
        Execute the one-round update in Algorithm 2 and **continue**
    Compute $U_t(x_t, p) = +\infty$ if $W_t(x_t, p) = 0$, and otherwise use

$$U_t(x_t, p) = \widehat{G}_t^{\mathrm{cal}}(x_t, p) + C_{\mathrm{tune}}\sqrt{\frac{\log(T|\Gamma_{\mathrm{dec}}|/h)}{\max\{1, W_t(x_t, p)\}}} + C_{\mathrm{tune}}(\varepsilon + h^\alpha)$$

    Choose $p_t \in \arg\max_{p \in \Gamma_{\mathrm{dec}}} U_t(x_t, p)$, let $a_t = \pi(p_t)$ and $y_t = \widetilde{y}_z(a_t)$, play (p_t, y_t), and observe O_t
    Append (x_t, O_t, Y_t), where $Y_t = (p_t + b)O_t - h_c(y_t - O_t)$, to the history of p_t

The startup fallback covers horizons for which $h > h_{\mathrm{vis}}$, before quantile visibility is certified. For all sufficiently large T, the profiling–UCB policy is used. The finite remaining range of horizons is absorbed into the theorem constant. On UCB rounds, summing local confidence widths along the realized trajectory gives $\widetilde{O}(\sqrt{T|\Gamma|h^{-d}})$. Together with the profiling cost and approximation losses above, this leads to the regret bounds in Section 5. The concentration and counting arguments are given in Appendix E.

Algorithm 2 Inventory Search

Require: horizon T, coarse profiler partition $\mathcal{Z}_{\text{prof}}$, current context x_t, its cell $z = z_{\text{prof}}(x_t)$, persistent profiler state for z, value accuracy ε, anchor grid Γ_y, confidence level δ, fixed tuning constants C_{tune} and $c_\eta > 0$

At initialization, set $\eta = c_\eta\sqrt{\varepsilon}$ and

$$B_y = 1 + \left\lceil \log_2\left(1 + \frac{y_{\max} - y_{\min}}{C_{\text{tune}}\eta}\right)\right\rceil, \qquad m_y = \left\lceil C_{\text{tune}}\eta^{-2}\log(T|\Gamma_y||\mathcal{Z}_{\text{prof}}|B_y/\delta)\right\rceil$$

At initialization, for every $(z,a) \in \mathcal{Z}_{\text{prof}} \times \Gamma_y$, initialize $[L_{z,a}, U_{z,a}] = [y_{\min}, y_{\max}]$
Select the next unfinished anchor a in the round-robin schedule of cell z
Set $p_t = a$ and $y_t = (L_{z,a} + U_{z,a})/2$, play (p_t, y_t), observe O_t, and append $I_t = \mathbf{1}\{O_t \geq y_t\}$ to the current batch of (z,a)
if the current batch contains m_y indicators **then**
    Let $\widehat{S}$ be their average
    if $\widehat{S} > \tau(a) + 2\eta$ **then**
        Set $L_{z,a} = y_t$
    else if $\widehat{S} < \tau(a) - 2\eta$ **then**
        Set $U_{z,a} = y_t$
    else
        Mark (z,a) finished and set $\widetilde{y}_z(a) = y_t$
    if (z,a) is unfinished and $U_{z,a} - L_{z,a} \leq C_{\text{tune}}\eta$ **then**
        Mark (z,a) finished and set $\widetilde{y}_z(a) = (L_{z,a} + U_{z,a})/2$
    Clear the completed batch
if all (z,a'), $a' \in \Gamma_y$, are finished **then**
    Mark cell z profiled and define $\widetilde{y}_x(p) = \widetilde{y}_z(\pi(p))$ for $x \in C_z^{\text{prof}}$, $p \in \Gamma_{\text{dec}}$, where $\pi(p) \in \arg\min_{a\in\Gamma_y}|p-a|$

C Formal Statements and Tuning

C.1 Tuning Convention

The policy inputs are the horizon T, context smoothness exponent α, calibration level ρ, price-regularity mode, and the fixed tuning constants $C_{\text{tune}}, h_{\text{vis}}, c_\eta$. The action ranges and costs are given, and the kernel is chosen as part of the policy. For the theoretical guarantees, tuning uses known numerical bounds for the model class. These include bounds on

$$L_x, L_p, L_F, L_{\text{loc}}, K, C_y, Y_{\max}, R_{\max}, c_y^{-1}, \kappa^{-1}, r_y^{-1}, r_q^{-1}, m_q^{-1}$$

and fixed action/kernel constants. Here $K := \kappa'$ is the density upper bound, and $Y_{\max}, R_{\max}$ bound proxy magnitude and one-period regret. Bounds for derived quantities are obtained from the primitive model bounds. The smooth-price class also fixes a bound $L_{G,2}$. Fix a finite multiplier $c_\eta > 0$, independent of T, and set $\eta = c_\eta\sqrt{\varepsilon}$. The choice $c_\eta = 1$ recovers the unit-multiplier policy. Choose C_{tune} large enough for the stockout concentration bound and the stochastic and context-bias constants C_{ker} in Lemma 9, including the fixed factor comparing the logarithms in that lemma and the index. These requirements depend only on the numerical envelopes, not on the inventory approximation error produced by the stopping rule.

To account for that error separately, define

$$\begin{aligned} L_\phi &:= \sup_{p\in\mathcal{P}}|\phi'(p)| = \frac{h_c}{(p_{\min} + b + h_c)^2}, \\ C_{\text{prof}} &:= 1 + L_{\text{loc}} + L_p + L_\phi/\kappa + c_\eta\max\{C_{\text{tune}}/2, 3/\kappa\}, \\ B_{\text{prof}} &:= C_y C_{\text{prof}}^2. \end{aligned}$$

After fixing C_{tune} and c_η, choose $0 < h_{\mathrm{vis}} \le 1$ so small that, for every $h \le h_{\mathrm{vis}}$,

$$\begin{aligned} C_{\mathrm{prof}}\sqrt{h^\alpha} &\le \min\{r_y, r_q/4, m_q/2\}, \\ L_F h^\alpha &\le \kappa r_q/2, \\ (L_x + L_F/\kappa)h^\alpha &\le \min\{m_q/8, r_q/4\}, \\ C_{\mathrm{tune}} h^\alpha &\le \tfrac{1}{4}\sqrt{h^\alpha}. \end{aligned}$$

Such a choice exists because $\alpha > 0$. The constants $C_{\mathrm{prof}}, B_{\mathrm{prof}}$ quantify the inventory approximation error in the proof. The upper-bound analysis retains this error through a common analytical shift of the indices. Fixed admissible tuning choices affect multiplicative constants and finite-horizon transients. The regret exponent is unchanged. Upper-bound constants may depend on c_η and its inverse. The demand surface and residual laws remain unknown.

C.2 Auxiliary Statements

Lemma 2 (Unique optimal inventory). *Under Assumptions 1 and 3, for every context x and price p, $y_x^*(p) = \lambda(x,p) + z_p(x)$ is the unique maximizer of $Q_x(p,\cdot)$ over $[y_{\min}, y_{\max}]$. It is also the unique solution of $S_x(y \mid p) = \tau(p)$.*

Lemma 3 (Quantile separation). *Under Assumptions 1 and 3, the calibration and critical quantiles satisfy, uniformly in x, p,*

$$y_x^*(p) - q_\rho(x,p) = z_p(x) - z_\rho(x) \ge \frac{\inf_{q\in\mathcal{P}} \phi(q) - \rho}{\kappa'} > 0.$$

Throughout the analysis, m_q denotes a fixed positive constant smaller than this lower bound.

Lemma 4 (Lipschitz price values). *Under Assumptions 1 and 2, there is a finite L_G, depending only on the model bounds, such that*

$$|G_x(p) - G_x(q)| \le L_G|p-q| \qquad \text{for all } x \in \mathcal{X},\ p,q \in \mathcal{P}.$$

Proposition 3 (Location-model separation). *Under Assumptions 1 to 3, $y_x^*(p)$ is the unique maximizer of $Q_x(p,\cdot)$, and*

$$G_x(p) = p\lambda(x,p) + C_x(p),$$

where

$$C_x(p) = p\mathbb{E}_{F_x}[\min\{\epsilon, z_p(x)\}] - b\mathbb{E}_{F_x}[(\epsilon - z_p(x))^+] - h_c\mathbb{E}_{F_x}[(z_p(x) - \epsilon)^+].$$

Moreover, there is a fixed $L_{\mathrm{loc}} < \infty$ such that, uniformly in p, each of $y_x^*(p)$, $q_\rho(x,p)$, $G_x(p)$, and $\overline{G}_x(p)$ changes by at most $L_{\mathrm{loc}}\|x - x'\|^\alpha$ between contexts x, x'.

Proposition 4 (Uniform price discretization). *Under Assumptions 1 and 2,*

$$|\Gamma(\varepsilon)| \le 2 + \frac{p_{\max} - p_{\min}}{\varepsilon}$$

and

$$\sup_x \left\{ \max_{p\in\mathcal{P}} G_x(p) - \max_{q\in\Gamma(\varepsilon)} G_x(q) \right\} \le L_G\varepsilon.$$

Under Assumption 4, the grid $\Gamma(\sqrt{\varepsilon})$ has cardinality $O(\varepsilon^{-1/2})$ and approximation error $O(\varepsilon)$.

Proposition 5 (Preservation of inventory profit gaps). *The observable surrogate preserves all expected profit differences between inventory decisions at a fixed context and price. Specifically, for every x, p and any feasible y, y',*

$$\overline{Q}_x(p,y) - \overline{Q}_x(p,y') = Q_x(p,y) - Q_x(p,y').$$

Consequently, it preserves both the optimal inventory and the loss from choosing a suboptimal inventory:

$$\arg\max_y \overline{Q}_x(p,y) = \arg\max_y Q_x(p,y), \qquad \overline{G}_x(p) - \overline{Q}_x(p,y) = G_x(p) - Q_x(p,y).$$

Lemma 5 (Mean calibration up to a within-context common shift)**.** *Under Assumption 1, define* $q_\rho(x,p) = \lambda(x,p) + z_\rho(x)$*. Then*

$$\overline{G}_x(p) - bq_\rho(x,p) = G_x(p) - bz_\rho(x).$$

For a fixed x*, the rightmost offset is independent of* p *and* y*. Hence an upper confidence bound for the left-hand side orders prices exactly as an upper confidence bound for* $G_x(p)$*.*

Lemma 6 (Two-resolution asynchronous profiling guarantee)**.** *Under Assumptions 1 to 3, take* $\varepsilon = h^\alpha$ *and* $\eta = c_\eta\sqrt{\varepsilon}$ *for a fixed* $c_\eta > 0$*, with tuning as in Appendix C.1. The constant* C *below may depend on* c_η*. Let* $N_{\mathrm{prof}}(T)$ *be the number of rounds up to horizon* T *on which the asynchronous policy performs a profiling update. Deterministically,*

$$N_{\mathrm{prof}}(T) \le C|\Gamma_y|\delta_x^{-d}\varepsilon^{-1}\log^2\left(\frac{CT|\Gamma_y|\delta_x^{-d}}{\delta}\right)$$

and, for $h \le h_{\mathrm{vis}}$*, with probability at least* $1-\delta$*, every completed cell has returned estimates satisfying, uniformly in* $p \in \Gamma_{\mathrm{dec}}$ *and* x *in that cell,*

$$|\widetilde{y}_x(p) - y_x^*(p)| \le C_{\mathrm{prof}}\sqrt{\varepsilon}$$

and

$$G_x(p) - Q_x(p,\widetilde{y}_x(p)) = \overline{G}_x(p) - \overline{Q}_x(p,\widetilde{y}_x(p)) \le B_{\mathrm{prof}}\varepsilon.$$

Under the certified resolution $h \le h_{\mathrm{vis}}$*,*

$$\widetilde{y}_x(p) \ge q_\rho(x,p) + m_q/2,$$

so the ρ*-quantile of* $\min\{D, \widetilde{y}_x(p)\}$ *equals* $q_\rho(x,p)$*. No completion guarantee for cells that have not received enough arrivals is needed. Since* $|\Gamma_y| = O(\varepsilon^{-1/2})$ *and* $\delta_x^{-d} = h^{-d/2}$*,*

$$N_{\mathrm{prof}}(T) \le \widetilde{O}\left(h^{-d/2}\varepsilon^{-3/2}\right).$$

C.3 Upper and Lower Bounds

Theorem 3 (Lipschitz-price contextual upper bound)**.** *Under Assumptions 1 to 3, with* $s = \mathrm{Lip}$*, and with the certified tuning inputs in Appendix C.1,* SP-*UCB satisfies*

$$\mathsf{Reg}_T \le CT^{\frac{d+2\alpha}{d+3\alpha}}\,\mathrm{polylog}(T).$$

Corollary 1 (Twice-smooth price upper bound)**.** *Under Assumptions 1 to 4, invoke* SP-*UCB with* $s = \mathrm{smooth}$*, using the same fixed tuning constants and fallback rules. Its price mesh is* $\Delta_p = \sqrt{\varepsilon}$*, with*

$$h = (\log(eT)/T)^{2/(2d+5\alpha)}, \qquad \varepsilon = h^\alpha.$$

Then

$$\mathsf{Reg}_T \le CT^{\frac{2d+3\alpha}{2d+5\alpha}}\,\mathrm{polylog}(T).$$

Theorem 4 (Context-conditioned shared-noise lower bounds)**.** *Assume* $b > 0$ *and* $h_c > 0$*, and let* $0 < \alpha \le 1$*. There exist fixed nondegenerate compact price and inventory ranges and a context-conditioned shared-noise location subclass satisfying Assumptions 1 to 3 such that every policy obeys, under Lipschitz price regularity,*

$$\sup_{\mathcal{I}} \mathsf{Reg}_T(\pi;\mathcal{I}) \ge cT^{\frac{d+2\alpha}{d+3\alpha}}.$$

For the subclass also satisfying Assumption 4, the lower bound is

$$cT^{\frac{2d+3\alpha}{2d+5\alpha}}.$$

Corollary 2 (Context-conditioned shared-noise minimax rates)**.** *Fix common numerical envelopes for all constants in Assumptions 1 to 3 and the fixed tuning constants, as well as fixed action ranges containing the subclass in Theorem 4. Let* $\mathfrak{I}_{\mathrm{Lip}}$ *be the resulting context-conditioned shared-noise class with Lipschitz price values, and let* $\mathfrak{I}_{\mathrm{sm}}$ *additionally satisfy Assumption 4 with a fixed bound* $L_{G,2}$*. For fixed* $b, h_c > 0$*, the upper and lower bounds yield*

$$\inf_\pi \sup_{\mathcal{I}\in\mathfrak{I}_{\mathrm{Lip}}} \mathsf{Reg}_T(\pi;\mathcal{I}) = \widetilde{\Theta}\left(T^{\frac{d+2\alpha}{d+3\alpha}}\right).$$

Under twice-smooth price values the corresponding rate is

$$\inf_\pi \sup_{\mathcal{I}\in\mathfrak{I}_{\mathrm{sm}}} \mathsf{Reg}_T(\pi;\mathcal{I}) = \widetilde{\Theta}\left(T^{\frac{2d+3\alpha}{2d+5\alpha}}\right).$$

D Proofs for the Proxy and Profiling Results

D.1 Profit Identities and Inventory Properties

Proof of Proposition 5. The elementary identity

$$D = \min\{D, y\} + (D - y)^+$$

gives

$$\begin{aligned} r((p, y), D) + bD &= pO - b(D - y)^+ - h_c(y - D)^+ + b\{O + (D - y)^+\} \\ &= (p + b)O - h_c(y - O), \end{aligned}$$

because $y - O = (y - D)^+$. This is exactly $Y(p, y, O)$. Taking conditional expectations and using mean-zero noise yields

$$\overline{Q}_x(p, y) = Q_x(p, y) + b\lambda(x, p).$$

For fixed x, p, the added term is independent of y. Subtracting the identities at y and y' proves the claimed equality of profit differences and hence the equality of maximizers. Maximizing over inventory also gives

$$\overline{G}_x(p) = G_x(p) + b\lambda(x, p).$$

Subtracting the preceding identity for $\overline{Q}_x(p, y)$ proves the equality of suboptimality losses. □

Proof of Lemma 5. The location family gives $q_\rho(x, p) = \lambda(x, p) + z_\rho(x)$. Substitute this identity into $\overline{G}_x(p) = G_x(p) + b\lambda(x, p)$. □

Proof of Lemma 2. Fix x, p, abbreviate $\lambda = \lambda(x, p)$, and write $y_x^* = y_x^*(p) = \lambda + z_p(x)$. Since $F_x(z_p(x)) = \phi(p)$,

$$S_x(y_x^* \mid p) = 1 - F_x(z_p(x)) = \tau(p).$$

For $y < y_x^*$, monotonicity gives $S_x(y \mid p) \geq \tau(p)$, while for $y > y_x^*$ it gives $S_x(y \mid p) \leq \tau(p)$. The lower density bound around $z_p(x)$ makes these inequalities strict away from y_x^*, so the root is unique. Because $y_x^*(p)$ is feasible with a margin by Assumption 3, the first-order identity in Section 4, together with the inventory-independent surrogate offset, implies that it is also the unique maximizer of $Q_x(p, \cdot)$. □

Proof of Lemma 3. By Lemma 2 and the definition of the calibration quantile, $y_x^*(p) - q_\rho(x, p) = z_p(x) - z_\rho(x)$. The density upper bound gives

$$\phi(p) - \rho = \int_{z_\rho(x)}^{z_p(x)} f_x(u)\, du \leq \kappa'\{z_p(x) - z_\rho(x)\}.$$

Taking the infimum of $\phi(p)$ over the price range proves the claim. □

Proof of Lemma 1. Fix x, p, and write $\lambda = \lambda(x, p)$ and $y_x^* = y_x^*(p) = \lambda + z_p(x)$. For $|y - y_x^*| \leq r_y \leq r_q$,

$$\begin{aligned} |S_x(y \mid p) - \tau(p)| &= |F_x(y - \lambda) - F_x(z_p(x))| \\ &= \left| \int_{z_p(x)}^{y-\lambda} f_x(u)\, du \right|. \end{aligned}$$

The density bounds in Assumption 3 therefore give

$$\kappa|y - y_x^*| \leq |S_x(y \mid p) - \tau(p)| \leq K|y - y_x^*|.$$

Finally, integrating $\partial_y Q_x(p, y) = (p + b + h_c)\{S_x(y \mid p) - \tau(p)\}$ between y and y_x^* yields

$$0 \leq G_x(p) - Q_x(p, y) \leq \frac{p_{\max} + b + h_c}{2} K|y - y_x^*|^2.$$

The claim follows after taking $c_y = \kappa$ and increasing C_y if necessary. □

D.2 Regularity and Approximation

Proof of Proposition 3. Fix x, p, abbreviate $\lambda = \lambda(x,p)$, and write $y_x^*(p) = \lambda + z_p(x)$. By Lemma 2, this inventory is the unique maximizer of $Q_x(p,\cdot)$.

At this inventory, the following identities hold sample by sample:

$$\begin{aligned}\min\{\lambda(x,p)+\epsilon, \lambda(x,p)+z_p(x)\} &= \lambda(x,p)+\min\{\epsilon, z_p(x)\},\\ (\lambda(x,p)+\epsilon-\lambda(x,p)-z_p(x))^+ &= (\epsilon - z_p(x))^+,\\ (\lambda(x,p)+z_p(x)-\lambda(x,p)-\epsilon)^+ &= (z_p(x)-\epsilon)^+.\end{aligned}$$

Substitution into the definition of Q_x gives

$$\begin{aligned}G_x(p) &= p\lambda(x,p) + p\mathbb{E}_{F_x}[\min\{\epsilon, z_p(x)\}] - b\mathbb{E}_{F_x}[(\epsilon - z_p(x))^+] - h_c\mathbb{E}_{F_x}[(z_p(x)-\epsilon)^+]\\ &= p\lambda(x,p) + C_x(p).\end{aligned}$$

It remains to prove the asserted context regularity. Put $d_{xx'} = \|x - x'\|$. For $u \in \{\rho\} \cup \{\phi(p) : p \in \mathcal{P}\}$, put $\delta_{xx'} = L_F d_{xx'}^\alpha$. When $\delta_{xx'} < \kappa r_q$, the two quantiles must be within r_q: were, say, $z_u(x') > z_u(x) + r_q$, integrating $f_x \geq \kappa$ from $z_u(x)$ to $z_u(x) + r_q$ would give $F_{x'}(z_u(x)+r_q) \geq F_x(z_u(x)+r_q) - \delta_{xx'} > u$, contradicting the definition of $z_u(x')$. Inside that neighborhood, the same integration argument gives

$$\kappa|z_u(x) - z_u(x')| \leq \delta_{xx'}.$$

When $\delta_{xx'} \geq \kappa r_q$, compact support gives

$$|z_u(x) - z_u(x')| \leq \overline{z} - \underline{z} \leq \frac{(\overline{z}-\underline{z})L_F}{\kappa r_q} d_{xx'}^\alpha.$$

Thus both $z_\rho(x)$ and $z_p(x)$ are uniformly Hölder in x, and the same is true of $q_\rho(x,p)$ and $y_x^*(p)$ after adding the location bound from Assumption 2.

For completeness, let $P_{x,p}$ denote the law of $\lambda(x,p)+\epsilon$. In one dimension,

$$\begin{aligned}W_1(P_{x,p}, P_{x',p}) &\leq |\lambda(x,p) - \lambda(x',p)| + W_1(F_x, F_{x'})\\ &\leq \{L_x + (\overline{z}-\underline{z})L_F\} d_{xx'}^\alpha,\end{aligned}$$

where the last step uses $W_1(F_x, F_{x'}) = \int_{\mathbb{R}} |F_x(u) - F_{x'}(u)|\, du$. For fixed (p,y), the profit is Lipschitz in demand with constant at most $L_r := \max\{p_{\max} + h_c, b\}$. The Kantorovich–Rubinstein inequality therefore bounds $|Q_x(p,y) - Q_{x'}(p,y)|$ by a constant times $d_{xx'}^\alpha$, uniformly in y. Taking maxima over the common inventory interval proves the same bound for $G_x(p)$. Finally, $\overline{G}_x(p) = G_x(p) + b\lambda(x,p)$, so $\overline{G}_x(p)$ is Hölder as well. Choosing one constant that dominates these four bounds proves the last claim. □

Proof of Lemma 4. Fix a context x, and couple demand at prices p, q using the same residual $\epsilon \sim F_x$. Then $D_p - D_q = \lambda(x,p) - \lambda(x,q)$. Let M be a common bound on demand and the absolute inventory levels, which exists by Assumption 1 and the compact action set. For fixed price and inventory, profit is Lipschitz in demand with constant $L_r = \max\{p_{\max} + h_c, b\}$. At a fixed demand realization, its price coefficient is $\min\{D, y\}$, whose absolute value is at most M. Therefore, for every feasible y,

$$\begin{aligned}|Q_x(p,y) - Q_x(q,y)| &\leq M|p-q| + L_r|\lambda(x,p) - \lambda(x,q)|\\ &\leq (M + L_r L_p)|p-q|.\end{aligned}$$

Taking maxima over the common inventory interval yields the claimed bound with $L_G = M + L_r L_p$. □

Proof of Proposition 4. The displayed cardinality bound follows directly from the construction of $\Gamma(\varepsilon)$. Fix a maximizer $p_x^* \in \arg\max_{p\in\mathcal{P}} G_x(p)$. It has a grid point within distance ε; the Lipschitz bound then gives

$$G_x(p_x^*) - \max_{q\in\Gamma(\varepsilon)} G_x(q) \leq L_G \varepsilon.$$

Under the bounded-second-derivative condition, Taylor's theorem around an interior maximizer gives a loss at most $\frac{1}{2}L_{G,2}|p - p_x^*|^2$. The same order follows at a boundary maximizer because both endpoints themselves belong to the grid, and hence the discretization loss is then zero. A grid of mesh $\sqrt{\varepsilon}$ therefore incurs $O(\varepsilon)$ loss and contains $O(\varepsilon^{-1/2})$ points. □

Lemma 7 (Within-cell approximation)**.** *For a sufficiently fine context grid, the inventory $\overline{y}_C(p)$ maximizing the cell-average expected surrogate score is unique and satisfies*

$$\sup_{x \in C} |\overline{y}_C(p) - y_x^*(p)| \le L_{\mathrm{loc}} \delta_x^{\alpha}$$

for every cell C of positive probability and every price p.

Context pooling and estimation error. The resolution condition in Lemma 7 is satisfied by $L_{\mathrm{loc}}\delta_x^{\alpha} \le r_q/2$. The cell inventory target is characterized by

$$\overline{S}_C(\overline{y}_C(p) \mid p) = \tau(p), \qquad \overline{S}_C(y \mid p) := \mathbb{E}[S_X(y \mid p) \mid X \in C].$$

Any feasible estimate $\widehat{y}_C(p)$ satisfying $|\widehat{y}_C(p) - \overline{y}_C(p)| \le e$ obeys

$$\sup_{x \in C} |\widehat{y}_C(p) - y_x^*(p)| \le e + L_{\mathrm{loc}}\delta_x^{\alpha}.$$

Whenever $e + L_{\mathrm{loc}}\delta_x^{\alpha} \le r_y$, it also satisfies

$$\sup_{x \in C} \big(G_x(p) - Q_x(p, \widehat{y}_C(p))\big) \le C_y \big(e + L_{\mathrm{loc}}\delta_x^{\alpha}\big)^2.$$

Proof of Lemma 7 and the estimation bounds. Fix a positive-probability cell C and a price p. By Proposition 3,

$$|y_x^*(p) - y_{x'}^*(p)| \le L_{\mathrm{loc}}\delta_x^{\alpha} \qquad (x, x' \in C).$$

Put $\ell = \inf_{x \in C} y_x^*(p)$ and $u = \sup_{x \in C} y_x^*(p)$. The inventory range assumption makes both endpoints feasible. Continuity and monotonicity give

$$\overline{S}_C(\ell \mid p) \ge \tau(p), \qquad \overline{S}_C(u \mid p) \le \tau(p),$$

so there is a root $\overline{y}_C(p) \in [\ell, u]$. Since $u - \ell \le L_{\mathrm{loc}}\delta_x^{\alpha} \le r_q/2$, every point within $r_q/2$ of this root is within r_q of each $y_x^*(p)$. The density lower bound in Assumption 3 therefore makes $\overline{S}_C(\cdot \mid p)$ strictly decreasing in this neighborhood. Monotonicity then ensures global uniqueness.

The derivative of the cell-average expected surrogate score is

$$\partial_y \mathbb{E}[\overline{Q}_X(p, y) \mid X \in C] = (p + b + h_c)\big(\overline{S}_C(y \mid p) - \tau(p)\big).$$

Differentiation under the expectation is justified by the uniform bound on the inventory derivative. Its sign is positive below the unique root and negative above it, so the root uniquely maximizes this score. The bound $u - \ell \le L_{\mathrm{loc}}\delta_x^{\alpha}$ proves Lemma 7. The estimation bound follows by the triangle inequality, and the profit-loss bound follows from Lemma 1. □

D.3 Inventory Profiling Guarantees

Proof of Lemma 6. Put $\eta = c_\eta\sqrt{\varepsilon}$ for fixed $c_\eta > 0$, and $\delta_x = \varepsilon^{1/(2\alpha)}$. We first construct a single event on which every anchor-price bisection decision is correct. Fix a profiler cell C_z^{prof} with positive context probability and an anchor $a \in \Gamma_y$. Define its conditional-average stockout curve

$$\overline{S}_z(y \mid a) := \mathbb{E}[S_{x_t}(y \mid a) \mid x_t \in C_z^{\mathrm{prof}}].$$

Cells of zero probability receive no arrivals and contribute no profiling rounds. Every individual root in the cell equals $y_x^*(a) = \lambda(x, a) + z_a(x)$, and therefore

$$\sup_{x, x' \in C_z^{\mathrm{prof}}} |y_x^*(a) - y_{x'}^*(a)| \le L_{\mathrm{loc}}\delta_x^{\alpha} = L_{\mathrm{loc}}\sqrt{\varepsilon}.$$

Continuity and monotonicity place a root $\overline{y}_z^*(a)$ of $\overline{S}_z(\cdot \mid a) = \tau(a)$ between the smallest and largest individual roots. Consequently

$$\sup_{x \in C_z^{\mathrm{prof}}} |\overline{y}_z^*(a) - y_x^*(a)| \le L_{\mathrm{loc}}\sqrt{\varepsilon}. \tag{D.1}$$

By the tuning conditions, $L_{\mathrm{loc}}\sqrt{\varepsilon} \le r_q/4$. Every threshold within $r_q/2$ of $\overline{y}^*_z(a)$ is therefore within $3r_q/4$ of every individual root in the cell. Integrating the density bounds and then averaging shows that, for any two such thresholds $u < v$,

$$\kappa(v-u) \le \overline{S}_z(u \mid a) - \overline{S}_z(v \mid a) \le K(v-u).$$

Thus the averaged root is unique there. Monotonicity makes it unique globally.

The within-cell scheduler uses only the cell index and the past to schedule each update. Conditional on an arrival to the cell, the context therefore retains the fixed conditional law used in $\overline{S}_z$. During a fixed threshold batch, each stockout indicator is a bounded martingale difference after subtracting $\overline{S}_z(y \mid a)$. Hoeffding's inequality gives stochastic error at most η from

$$m_y \ge C\eta^{-2}\log\left(\frac{|\Gamma_y||\mathcal{Z}_{\mathrm{prof}}|B_y}{\delta}\right)$$

samples, where

$$B_y := 1 + \left\lceil \log_2\left(1 + \frac{y_{\max} - y_{\min}}{C_{\mathrm{tune}}\eta}\right)\right\rceil$$

bounds the number of thresholds queried by one bisection. Taking the union bound over the at most $|\Gamma_y||\mathcal{Z}_{\mathrm{prof}}|B_y$ adaptive batches yields an event $\mathcal{E}_{\mathrm{bis}}$ of probability at least $1 - \delta/2$ on which

$$|\widehat{S} - \overline{S}_z(y \mid a)| \le \eta \tag{D.2}$$

for every queried batch. The union bound is valid for adaptive thresholds because the conditional tail bound holds at every batch given its starting history. A batch unfinished at T can be continued on subsequent cell arrivals for this argument. Positive cell probability ensures that the extended batch finishes almost surely. Bounding the extended batch also bounds the event that an actually completed batch is inaccurate, without conditioning on completion by T.

We next verify the bisection invariant. Initially $\overline{y}^*_z(a) \in [L_{z,a}, U_{z,a}]$. The averaged stockout curve is nonincreasing. On $\mathcal{E}_{\mathrm{bis}}$, $\widehat{S} > \tau(a) + 2\eta$ implies $\overline{S}_z(y \mid a) > \tau(a) + \eta$, hence $y < \overline{y}^*_z(a)$ and replacing the lower endpoint by y preserves the invariant. The symmetric statement holds when $\widehat{S} < \tau(a) - 2\eta$. In the remaining case,

$$|\overline{S}_z(y \mid a) - \tau(a)| \le 3\eta.$$

The choice of C_{prof} and h_{vis} gives $3\eta \le \kappa r_q/4$. A threshold at distance at least $r_q/2$ from the averaged root has stockout gap at least $\kappa r_q/2$, by the preceding slope bound and monotonicity. Hence the ambiguous branch terminates inside that neighborhood, where the same bound gives

$$|y - \overline{y}^*_z(a)| \le 3\eta/\kappa. \tag{D.3}$$

A width-based termination has error at most $C_{\mathrm{tune}}\eta/2$, since its bracket still contains the root. Consequently every cell–anchor output satisfies

$$|\widetilde{y}_z(a) - \overline{y}^*_z(a)| \le \max\{C_{\mathrm{tune}}/2, 3/\kappa\}\eta. \tag{D.4}$$

We now transfer the anchor estimate to a decision price $p \in \Gamma_{\mathrm{dec}}$. Since ϕ has bounded derivative and $f_x \ge \kappa$ along the quantile curve $\{z_q(x) : q \in \mathcal{P}\}$,

$$\kappa|z_p(x) - z_a(x)| \le |F_x(z_p(x)) - F_x(z_a(x))| = |\phi(p) - \phi(a)| \le C|p-a|.$$

Here the first inequality follows by integrating the density between the two quantiles. Every intermediate quantile is $z_q(x)$ for a price between p and a. Thus $p \mapsto z_p(x)$ is uniformly Lipschitz under the context-conditioned shared-noise assumption. With $a = \pi(p)$, $|p-a| \le \sqrt{\varepsilon}$. Since Lemma 2 gives $y^*_x(p) = \lambda(x,p) + z_p(x)$, for $x \in C^{\mathrm{prof}}_z$,

$$\begin{aligned}
|\widetilde{y}_z(a) - y^*_x(p)| &\le |\widetilde{y}_z(a) - \overline{y}^*_z(a)| + |\overline{y}^*_z(a) - y^*_x(a)| + |\lambda(x,a) - \lambda(x,p)| + |z_a(x) - z_p(x)| \\
&\le \max\{C_{\mathrm{tune}}/2, 3/\kappa\}\eta + L_{\mathrm{loc}}\sqrt{\varepsilon} + (L_p + L_\phi/\kappa)|p-a| \\
&\le C_{\mathrm{prof}}\sqrt{\varepsilon}.
\end{aligned} \tag{D.5}$$

The error constant is therefore

$$C_{\mathrm{prof}} = 1 + L_{\mathrm{loc}} + L_p + L_\phi/\kappa + c_\eta \max\{C_{\mathrm{tune}}/2, 3/\kappa\}.$$

Only the last term is reduced by decreasing c_η: the cell and anchor approximation errors retain resolution $\sqrt{\varepsilon}$.

It remains to count the rounds that actually perform profiling. Writing $M = |\Gamma_y||\mathcal{Z}_{\text{prof}}|$, the exact deterministic cap is

$$N_{\text{prof}}(T) \le \min\{T, MB_y m_y\}, \qquad m_y = \left\lceil C_{\text{tune}} c_\eta^{-2} \varepsilon^{-1} \log(TMB_y/\delta) \right\rceil,$$

where B_y uses $C_{\text{tune}} c_\eta \sqrt{\varepsilon}$ in its denominator. Thus the batch cost scales as c_η^{-2} before rounding, with an additional logarithmic dependence through B_y. For each fixed positive c_η, these factors can be absorbed into constants in the following bounds, preserving the exponent of ε. Each cell–anchor pair uses at most B_y batches, and each batch uses m_y arrivals assigned to that cell. Hence every cell uses at most

$$R_{\text{prof}} \le C|\Gamma_y|\varepsilon^{-1} \log(e/\varepsilon) \log\left(\frac{T|\Gamma_y||\mathcal{Z}_{\text{prof}}|}{\delta}\right) \tag{D.6}$$

profiling actions. Since $|\mathcal{Z}_{\text{prof}}| \le C\delta_x^{-d}$, summing this deterministic cap over cells gives, for every arrival realization and every horizon,

$$N_{\text{prof}}(T) \le C|\Gamma_y|\delta_x^{-d}\varepsilon^{-1} \log^2\left(\frac{CT|\Gamma_y|\delta_x^{-d}}{\delta}\right). \tag{D.7}$$

An unfinished cell cannot create waiting loss elsewhere: each arrival to an unfinished cell is one of the actions counted above, and each arrival to a finished cell is a UCB round. On $\mathcal{E}_{\text{bis}}$, every cell that does finish has the accuracy guarantees in Eqs. (D.4) and (D.5). Cells that have not received enough arrivals remain in profiling mode and never supply an uncertified inventory to the UCB history.

Under $h \le h_{\text{vis}}$, the certified bound on Eq. (D.5) lies inside the radius in Assumption 3. The quadratic loss bound in Lemma 1 and $\varepsilon = h^\alpha \le 1$ give

$$G_x(p) - Q_x(p, \widetilde{y}_x(p)) \le C_y C_{\text{prof}}^2 \varepsilon = B_{\text{prof}}\varepsilon.$$

The equality of the proxy and true inventory gaps follows from Proposition 5. Finally, $y_x^*(p) - q_\rho(x,p) \ge m_q$, and the definition of h_{vis} makes the right-hand side of Eq. (D.5) at most $m_q/2$. Hence $\widetilde{y}_x(p) \ge q_\rho(x,p) + m_q/2$. At every $u \le q_\rho(x,p)$, censoring is therefore inactive, so the ρ-quantile of sales equals the ρ-quantile of demand. □

Proof of Proposition 1. Fix p and apply the bisection procedure with the singleton anchor set $\Gamma_y = \{p\}$. For sufficiently small ε, $L_{\text{loc}}\delta_x^\alpha \le L_{\text{loc}}\sqrt{\varepsilon} \le r_q/4$. By Lemma 7, each positive-probability cell has a unique target $\overline{y}_C(p)$, characterized by $\overline{S}_C(\overline{y}_C(p) \mid p) = \tau(p)$, and

$$\sup_{x \in C} |\overline{y}_C(p) - y_x^*(p)| \le L_{\text{loc}}\delta_x^\alpha.$$

The concentration and bisection-invariant argument in the proof of Lemma 6 applies unchanged to any partition with $\delta_x^\alpha \le \sqrt{\varepsilon}$. In particular, Eq. (D.4), with $\eta = c_\eta\sqrt{\varepsilon}$, gives, simultaneously for every completed cell, with probability at least $1-\delta$,

$$|\widehat{y}_C(p) - \overline{y}_C(p)| \le C_0\sqrt{\varepsilon}, \qquad C_0 := c_\eta \max\{C_{\text{tune}}/2, 3/\kappa\}.$$

Combining this bound with Lemma 7 by the triangle inequality gives

$$\sup_{x \in C} |\widehat{y}_C(p) - y_x^*(p)| \le (C_0 + L_{\text{loc}})\sqrt{\varepsilon} = C_{\text{inv}}\sqrt{\varepsilon}, \qquad C_{\text{inv}} := C_0 + L_{\text{loc}}.$$

This constant depends only on the model constants and fixed algorithm parameters, not on T, ε, or δ_x.

Each cell uses at most $B_y m_y = \widetilde{O}(\varepsilon^{-1})$ inventory-search rounds, as in Eq. (D.6) with one anchor. Each such round incurs regret at most $R_{\max}$, so the cumulative regret incurred during inventory search in each completed cell is at most

$$R_{\max} B_y m_y = \widetilde{O}(\varepsilon^{-1}).$$

The same upper bound applies to the search rounds already used in an unfinished cell. Zero-probability cells contribute no rounds almost surely. □

E Proofs for Kernel Concentration and Counting

Lemma 8 (Sequential kernel proxy and quantile concentration). *Let $\ell_T(\delta) = \log(CT^4|\Gamma|h^{-d}/\delta)$. There is an event $\mathcal{E}_{\mathrm{ker}}$ with probability at least $1-\delta$ such that, on its intersection with the profiling event, uniformly over UCB queries with $W_t(x,p) > 0$,*

$$|\widehat{\overline{G}}_t(x,p) - \overline{G}_x(p)| \le C\left(\sqrt{\frac{\ell_T(\delta)}{W_t(x,p)}} + \varepsilon + h^\alpha\right),$$

$$|\widehat{q}_{\rho,t}(x,p) - q_\rho(x,p)| \le C\left(\sqrt{\frac{\ell_T(\delta)}{W_t(x,p)}} + h^\alpha\right).$$

Lemma 9 (Kernel calibrated-value confidence). *On the intersection of the profiling event and $\mathcal{E}_{\mathrm{ker}}$, uniformly over UCB queries with $W_t(x,p) > 0$,*

$$\left|\widehat{G}_t^{\mathrm{cal}}(x,p) - \{G_x(p) - bz_\rho(x)\}\right| \le C\sqrt{\frac{\ell_T(\delta)}{\max\{1, W_t(x,p)\}}} + C(\varepsilon + h^\alpha).$$

More precisely, write $H_x(p) = G_x(p) - bz_\rho(x)$ and let

$$\mathfrak{b}_t(x,p) := \frac{\sum_{s\in\mathcal{T}_t(p)} K((x_s-x)/h)\{G_{x_s}(p) - Q_{x_s}(p,y_s)\}}{W_t(x,p)}.$$

Then $0 \le \mathfrak{b}_t(x,p) \le B_{\mathrm{prof}}\varepsilon$, and

$$\begin{aligned}\widehat{G}_t^{\mathrm{cal}}(x,p) - H_x(p) &= -\mathfrak{b}_t(x,p) + e_t^{\mathrm{cal}}(x,p),\\ |e_t^{\mathrm{cal}}(x,p)| &\le C_{\mathrm{ker}}\left\{\sqrt{\frac{\ell_T(\delta)}{\max\{1, W_t(x,p)\}}} + h^\alpha\right\}.\end{aligned} \tag{E.1}$$

Here C_{ker} depends only on the fixed model envelopes and not on C_{tune} or B_{prof}. The first, symmetric bound may also depend on the fixed B_{prof}.

Lemma 10 (Kernel counting). *Let $\mathcal{U}_T$ be the set of UCB rounds. For every predictable interleaving rule and price sequence $p_t \in \Gamma$ on those rounds,*

$$\sum_{t\in\mathcal{U}_T} \frac{1}{\sqrt{\max\{1, W_t(x_t,p_t)\}}} \le C\sqrt{T|\Gamma|h^{-d}}.$$

Proof of Lemma 8. We first make the sequential measurability explicit for the interleaved policy. Include the policy's internal random seed and initialized profiler state in $\mathcal{F}_0$, and let

$$\mathcal{F}_{s-1} = \sigma(\mathcal{F}_0, (x_r, p_r, y_r, O_r) : r < s), \qquad \mathcal{G}_s = \mathcal{F}_{s-1} \vee \sigma(x_s).$$

The decision whether round s is a profiling or UCB round, and then the action (p_s, y_s), are $\mathcal{G}_s$-measurable. By Assumption 1, $\epsilon_s \mid \mathcal{G}_s \sim F_{x_s}$. Write $A_s = 1$ when round s is a UCB round.

Fix a query time t, a price $p \in \Gamma$, and condition on $x_t = x$. For $s < t$, enlarge the pre-outcome sigma-field to

$$\mathcal{G}_s^{(t)} = \mathcal{G}_s \vee \sigma(x_t).$$

By the i.i.d. pair assumption and independence from policy randomization in Assumption 1, x_t is independent of the policy's random seed and all pairs (x_r, ϵ_r) with $r < t$. Thus $\epsilon_s \mid \mathcal{G}_s^{(t)} \sim F_{x_s}$, so adjoining x_t preserves the conditional law of the period-s outcome. Set

$$w_s = K((x_s - x)/h)\mathbf{1}\{A_s = 1, p_s = p\}, \qquad m_s^Y = \mathbb{E}[Y_s \mid \mathcal{G}_s].$$

The weight is $\mathcal{G}_s^{(t)}$-measurable, and $\mathbb{E}[Y_s - m_s^Y \mid \mathcal{G}_s^{(t)}] = 0$. Hence $w_s(Y_s - m_s^Y)$ is a bounded martingale difference in the filtration that reveals O_s after $\mathcal{G}_s^{(t)}$. Since $|Y_s - m_s^Y| \le 2Y_{\max}$ and $0 \le w_s \le 1$, Hoeffding's conditional lemma gives, for every $\theta \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left\{\theta w_s(Y_s - m_s^Y) - 2\theta^2 Y_{\max}^2 w_s^2\right\}\middle|\mathcal{G}_s^{(t)}\right] \le 1. \tag{E.2}$$

Thus the product of the factors in Eq. (E.2) is a nonnegative supermartingale. Also,

$$\sum_{s<t} w_s^2 \le \sum_{s<t} w_s = W_t(x,p).$$

For $W_t(x,p) > 1$, use the dyadic events $2^{j-1} < W_t(x,p) \le 2^j$, $j = 1, \dots, \lceil \log_2 T \rceil$. On a given level, the exponential bound with its deterministic variance cap 2^j gives

$$\mathbb{P}\left(\left| \sum_{s<t} w_s (Y_s - m_s^Y) \right| > C\sqrt{2^j L},\ W_t(x,p) \le 2^j \,\middle|\, x_t = x \right) \le 2e^{-L}.$$

Indeed, apply Markov's inequality to the exponential supermartingale and choose $|\theta|$ proportional to $\sqrt{L/2^j}/Y_{\max}$. Within this level $2^j < 2W_t(x,p)$. Assigning failure probability $\delta/(2CT|\Gamma|\log(eT))$ to each query and level therefore gives

$$\left| \sum_{s<t} w_s (Y_s - m_s^Y) \right| \le C\sqrt{W_t(x,p)\ell_T(\delta)} \tag{E.3}$$

for this conditional query. For $0 < W_t(x,p) \le 1$, boundedness gives the pathwise inequality

$$\left| \sum_{s<t} w_s (Y_s - m_s^Y) \right| \le 2Y_{\max} W_t(x,p) \le 2Y_{\max}\sqrt{W_t(x,p)\ell_T(\delta)}.$$

Thus Eq. (E.3) covers all positive weights. After division by $W_t(x,p) > 0$, the stochastic proxy error is at most $C\sqrt{\ell_T(\delta)/W_t(x,p)}$. This derivation handles the data-dependent effective sample size over its full positive range.

On the profiling event, every observation with $A_s = 1$ comes from a cell whose profile was already completed before choosing the action. Hence Proposition 5 and Lemma 6 give

$$|m_s^Y - \overline{G}_{x_s}(p)| = G_{x_s}(p) - Q_{x_s}(p, \widetilde{y}_{x_s}(p)) \le C\varepsilon. \tag{E.4}$$

Moreover, the context-regularity conclusion of Proposition 3 implies, whenever $w_s > 0$,

$$|\overline{G}_{x_s}(p) - \overline{G}_x(p)| \le L_{\mathrm{loc}} h^\alpha. \tag{E.5}$$

The weighted average of Eqs. (E.4) and (E.5), combined with Eq. (E.3), proves

$$|\widehat{\overline{G}}_t(x,p) - \overline{G}_x(p)| \le C\left\{ \sqrt{\frac{\ell_T(\delta)}{W_t(x,p)}} + \varepsilon + h^\alpha \right\}. \tag{E.6}$$

We now control the weighted empirical quantile. Put

$$r_\rho := \min\{r_q, m_q/4, r_y/2\}, \mathcal{U}_{x,p} = [q_\rho(x,p) - r_\rho, q_\rho(x,p) + r_\rho].$$

When $w_s > 0$, context smoothness gives

$$|q_\rho(x_s,p) - q_\rho(x,p)| \le (L_x + L_F/\kappa)h^\alpha.$$

Together with the profiling margin $\widetilde{y}_{x_s}(p) \ge q_\rho(x_s,p) + m_q/2$ and the certified inequality $(L_x + L_F/\kappa)h^\alpha \le m_q/8$, this shows that

$$u \le q_\rho(x,p) + r_\rho \le q_\rho(x_s,p) + r_\rho + m_q/8 < \widetilde{y}_{x_s}(p)$$

throughout $\mathcal{U}_{x,p}$. Thus censoring does not alter the lower-tail indicators:

$$\mathbf{1}\{O_s \le u\} = \mathbf{1}\{D_s \le u\}, \qquad u \in \mathcal{U}_{x,p}. \tag{E.7}$$

Let $\mathcal{V}_T$ be a deterministic grid of $[0, y_{\max}]$ with mesh at most T^{-2}, so $|\mathcal{V}_T| \le CT^2$. Conditional on $x_t = x$, augment the grid points lying in $\mathcal{U}_{x,p}$ by the two endpoints of that interval. This query-specific grid still has at most $CT^2 + 2$ points, all inside the uncensored neighborhood. Fix one such point v. Construct the concentration event using latent demand, before intersecting it with the profiling event. Specifically, enlarge the analysis filtration to

$$\mathcal{F}_{s-1}^D := \sigma(\mathcal{F}_0, (x_r, p_r, y_r, D_r) : r < s), \qquad \mathcal{J}_s^{(t)} := \mathcal{F}_{s-1}^D \vee \sigma(x_s, x_t).$$

Observed sales and all policy decisions are measurable in this larger analysis filtration. The policy continues to use its observed-sales history. For $s < t$, the i.i.d. pairs and independent policy randomization in Assumption 1 give $\epsilon_s \mid \mathcal{J}_s^{(t)} \sim F_{x_s}$, including after conditioning on latent past demand and the query context. Define

$$\xi_s^D(v) := \mathbf{1}\{D_s \le v\} - F_{x_s}(v - \lambda(x_s, p_s)).$$

The weights remain predictable and $\mathbb{E}[w_s \xi_s^D(v) \mid \mathcal{J}_s^{(t)}] = 0$ on the full probability space. Consequently $w_s \xi_s^D(v)$ obeys the same exponential-supermartingale argument as the proxy sequence, without any conditioning on profiling success. For $0 < W_t \le 1$, use the deterministic bound $|\sum_{s<t} w_s \xi_s^D(v)| \le W_t$. Allocate failure probability $\delta/(2CT|\Gamma||\mathcal{V}_T|\log(eT))$ to every (t, p, v) and dyadic weight level. A union bound gives the desired inequality for the latent weighted CDF at all relevant grid points. We then intersect this concentration event with the profiling event. On the intersection, Eq. (E.7) identifies the latent and observed indicators for every positive weight, and $p_s = p$ on those weights. Thus the unconditional concentration bound for latent demand transfers to the empirical sales CDF on the profiling event.

For an arbitrary $u \in \mathcal{U}_{x,p}$, take adjacent grid points $v_- \le u \le v_+$. Monotonicity gives $\widehat{F}_t(v_- \mid x, p) \le \widehat{F}_t(u \mid x, p) \le \widehat{F}_t(v_+ \mid x, p)$, while the population CDF changes by at most KT^{-2} between adjacent points. Absorbing this deterministic term into the radius yields, simultaneously over $u \in \mathcal{U}_{x,p}$,

$$\left| \widehat{F}_t(u \mid x, p) - \frac{\sum_{s<t} w_s F_{x_s}(u - \lambda(x_s, p))}{W_t(x, p)} \right| \le C \sqrt{\frac{\ell_T(\delta)}{W_t(x, p)}}. \tag{E.8}$$

The upper density bound, the Hölder condition on F_x, and $|\lambda(x_s, p) - \lambda(x, p)| \le L_x h^\alpha$ imply

$$\begin{aligned} &|F_{x_s}(u - \lambda(x_s, p)) - F_x(u - \lambda(x, p))| \\ &\quad \le \sup_v |F_{x_s}(v) - F_x(v)| + K|\lambda(x_s, p) - \lambda(x, p)| \\ &\quad \le (L_F + KL_x)h^\alpha. \end{aligned}$$

Consequently,

$$\sup_{u \in \mathcal{U}_{x,p}} |\widehat{F}_t(u \mid x, p) - F_x(u - \lambda(x, p))| \le e_t, \qquad e_t = C \left\{ \sqrt{\frac{\ell_T(\delta)}{W_t(x, p)}} + h^\alpha \right\}. \tag{E.9}$$

The target CDF equals ρ at $q_\rho(x, p)$. On the fixed neighborhood from Assumption 3, its density is at least κ. Suppose first that

$$\frac{2e_t}{\kappa} \le r_\rho. \tag{E.10}$$

Integrating the density lower bound on both sides of $z_\rho(x)$ gives

$$F_x(q_\rho(x, p) + 2e_t/\kappa - \lambda(x, p)) \ge \rho + 2e_t$$

and

$$F_x(q_\rho(x, p) - 2e_t/\kappa - \lambda(x, p)) \le \rho - 2e_t.$$

Both evaluation points lie in $\mathcal{U}_{x,p}$, so Eq. (E.9) implies that the empirical CDF is below ρ at the lower point and above ρ at the upper point. The defining crossing property of the measurable weighted empirical quantile then yields

$$|\widehat{q}_{\rho,t}(x, p) - q_\rho(x, p)| \le 2e_t/\kappa. \tag{E.11}$$

Outside the regime in Eq. (E.10),

$$e_t \ge e_0, \qquad e_0 := \frac{\kappa}{2} r_\rho > 0.$$

Both quantiles lie in the fixed bounded sales interval, so enlarging the universal radius constant by that interval's diameter divided by e_0 gives the same order Ce_t there as well.

The conditional failure bounds hold uniformly in $x_t = x$. Integrating over the query context gives the corresponding unconditional bounds. A union bound over at most $T|\Gamma|$ realized time–price queries, at most $1 + \lceil \log_2 T \rceil$ dyadic weight levels, and $|\mathcal{V}_T| \le CT^2$ CDF grid points has total failure probability at most δ. The logarithm of this cardinality is bounded by $\ell_T(\delta)$, which completes both simultaneous bounds over the realized queries. □

Proof of Lemma 9. Fix a query with positive weight and write $w_s = K((x_s - x)/h)\mathbf{1}\{s \in \mathcal{T}_t(p)\}$. The proxy identity gives $m_s^Y = \overline{G}_{x_s}(p) - \{G_{x_s}(p) - Q_{x_s}(p, y_s)\}$ whenever $w_s > 0$. Hence the following decomposition is exact:

$$\widehat{G}_t^{\mathrm{cal}}(x,p) - H_x(p) = -\mathfrak{b}_t(x,p) + \frac{\sum_{s<t} w_s(Y_s - m_s^Y)}{W_t(x,p)} + \frac{\sum_{s<t} w_s\{\overline{G}_{x_s}(p) - \overline{G}_x(p)\}}{W_t(x,p)} - b\{\widehat{q}_{\rho,t}(x,p) - q_\rho(x,p)\}.$$

The last three terms define e_t^{cal}. For $W_t \geq 1$, Eqs. (E.3) and (E.5) and Lemma 8 bound them by the right-hand side of Eq. (E.1), with a constant independent of the profiler's stopping tolerance. For $0 < W_t < 1$, the first of these terms is at most $2Y_{\max}$ in magnitude, the second is at most $L_{\mathrm{loc}}h^\alpha$, and the quantile difference is at most the fixed sales-range diameter. Since $\ell_T(\delta) \geq 1$, these deterministic bounds give the same result with denominator $\max\{1, W_t\}$. Finally, profiling bounds each inventory gap by $B_{\mathrm{prof}}\varepsilon$. Their weighted average lies in $[0, B_{\mathrm{prof}}\varepsilon]$. This proves the signed decomposition and, by the triangle inequality, the symmetric confidence bound. □

Proof of Lemma 10. Use a deterministic partition of $[0,1]^d$ into $M \leq Ch^{-d}$ cells of diameter at most $h/2$, and let $j(t)$ denote the cell containing x_t on a UCB round. Two contexts in the same cell are at distance at most $h/2$. Hence every earlier UCB observation in cell $j(t)$ at price p_t has kernel weight at least k_0. Let $N_{j,p}(t)$ be the number of such UCB observations strictly before t. Pathwise,

$$W_t(x_t, p_t) \geq k_0 N_{j,p_t}(t).$$

Order the visits to a fixed pair (j,p) by $r = 1, \ldots, N_{j,p}(T)$. On its r-th visit, the preceding count is $r-1$. The first visit contributes one because the denominator in the lemma is truncated at one, and

$$\sum_{r=1}^{N_{j,p}(T)} \frac{1}{\sqrt{\max\{1, r-1\}}} \leq C\sqrt{N_{j,p}(T)}.$$

After absorbing $k_0^{-1/2}$, summing this bound and applying Cauchy–Schwarz gives

$$\begin{aligned}\sum_{t \in \mathcal{U}_T} \frac{1}{\sqrt{\max\{1, W_t(x_t,p_t)\}}} &\leq C\sum_{j,p}\sqrt{N_{j,p}(T)} \\ &\leq C\sqrt{M|\Gamma|\sum_{j,p} N_{j,p}(T)} \\ &\leq C\sqrt{|\mathcal{U}_T||\Gamma|h^{-d}} \leq C\sqrt{T|\Gamma|h^{-d}}.\end{aligned}$$

This argument is deterministic once the context and price sequences are fixed, so it applies to every predictable adaptive interleaving policy. □

F Complete Proofs of the Contextual Upper Bounds

Complete proof of Theorem 3. Let $r = (d+2\alpha)/(d+3\alpha) < 1$ and $K_{\mathrm{dec}} = |\Gamma_{\mathrm{dec}}|$, $K_{\mathrm{prof}} = |\Gamma_y|$. The tuning constants are fixed problem constants. Since $h_T \to 0$ and

$$K_{\mathrm{prof}}\delta_x^{-d}\varepsilon^{-1}\,\mathrm{polylog}(T) = o(T),$$

there is a finite T_0 such that $h_T \leq h_{\mathrm{vis}}$ for every $T \geq T_0$. For $T < T_0$, every branch of Algorithm 1 has regret at most $R_{\max}T \leq R_{\max}T_0^{1-r}T^r$, so increasing the theorem constant covers these horizons. We henceforth take $T \geq T_0$. The deterministic startup fallback is then inactive.

Run the interleaved profiler–UCB policy with $\delta = T^{-2}$, and let $\mathcal{E}_{\mathrm{prof}}$ and $\mathcal{E}_{\mathrm{ker}}$ denote the events from Lemmas 6 and 8. For fixed d, α, the logarithm in the index is equivalent up to a constant to $\ell_T(T^{-2})$. The choice in Appendix C.1 makes the index radius dominate the C_{ker} bound on e_t^{cal}. The possibly larger inventory bias $B_{\mathrm{prof}}\varepsilon$ is kept separately below. A union bound gives

$$\mathbb{P}(\mathcal{E}^c) \leq 2T^{-2}, \qquad \mathcal{E} := \mathcal{E}_{\mathrm{prof}} \cap \mathcal{E}_{\mathrm{ker}}. \tag{F.1}$$

On $\mathcal{E}^c$, total regret is at most $TR_{\max}$, so its expected contribution is at most $2R_{\max}/T$.

For every arrival and outcome realization, Lemma 6 bounds the number of profiling rounds by

$$N_{\text{prof}}(T) \le CK_{\text{prof}}\delta_x^{-d}\varepsilon^{-1}\,\text{polylog}(T).$$

Every one of these rounds costs at most $R_{\max}$, so

$$\text{Reg}_{\text{prof}} \le CK_{\text{prof}}\delta_x^{-d}\varepsilon^{-1}\,\text{polylog}(T). \tag{F.2}$$

Here and below the constants may depend on the fixed c_η, including the batch factor c_η^{-2} and $B_{\text{prof}} = C_y C_{\text{prof}}^2$. For the stated bandwidth this quantity is $o(T)$. Arrivals to completed cells are UCB rounds and enter the UCB regret bound below.

Consider a UCB round with $x_t = x$. Let p_x^Γ maximize G_x over the grid. By Proposition 4,

$$V_x - G_x(p_x^\Gamma) \le L_G\varepsilon. \tag{F.3}$$

Recall $H_x(p) = G_x(p) - bz_\rho(x)$, and denote the actual finite index radius in Algorithm 1 by

$$\beta_t(x,p) := C_{\text{tune}}\sqrt{\frac{\log(T|\Gamma_{\text{dec}}|/h)}{\max\{1, W_t(x,p)\}}} + C_{\text{tune}}(\varepsilon + h^\alpha).$$

On $\mathcal{E}$, Eq. (E.1) and the tuning choice give $|e_t^{\text{cal}}(x,p)| \le \beta_t(x,p)$. Therefore, for every positive-weight price,

$$H_x(p) - B_{\text{prof}}\varepsilon \le U_t(x,p) \le H_x(p) + 2\beta_t(x,p). \tag{F.4}$$

The lower bound uses $\mathfrak{b}_t \le B_{\text{prof}}\varepsilon$. The upper bound uses $\mathfrak{b}_t \ge 0$. For analysis only, put $U_t^+(x,p) := U_t(x,p) + B_{\text{prof}}\varepsilon$, with $+\infty + B_{\text{prof}}\varepsilon = +\infty$. This added quantity is independent of the candidate price, so the policy's chosen maximizer is also a maximizer of U_t^+. The policy continues to use the original indices U_t. Whenever both relevant effective sample sizes are positive,

$$\begin{aligned} H_x(p_x^\Gamma) &\le U_t^+(x,p_x^\Gamma) \le U_t^+(x,p_t) \\ &\le H_x(p_t) + 2\beta_t(x,p_t) + B_{\text{prof}}\varepsilon. \end{aligned} \tag{F.5}$$

The first inequality follows from the lower confidence bound after the common shift, the second from the UCB selection rule, and the last from the upper confidence bound after the same shift. When $W_t(x,p_t) > 0$, a candidate with zero effective sample size has infinite index, so p_x^Γ cannot have zero weight. Hence the displayed chain indeed covers this case. When $W_t(x,p_t) = 0$, boundedness gives $G_x(p_x^\Gamma) - G_x(p_t) \le C$, which is absorbed by the same right-hand side because $\max\{1, W_t(x,p_t)\} = 1$ and $\ell_T(T^{-2}) \ge 1$. Since both B_{prof} and C_{tune} are fixed constants and the within-context shift cancels, Eq. (F.5) therefore yields for every UCB round

$$G_x(p_x^\Gamma) - G_x(p_t) \le C\sqrt{\frac{\ell_T(T^{-2})}{\max\{1, W_t(x,p_t)\}}} + C(\varepsilon + h^\alpha). \tag{F.6}$$

The period regret decomposes exactly as

$$\begin{aligned} V_x - Q_x(p_t, \widetilde{y}_x(p_t)) = \{V_x - G_x(p_x^\Gamma)\} + \{G_x(p_x^\Gamma) - G_x(p_t)\} \\ + \{G_x(p_t) - Q_x(p_t, \widetilde{y}_x(p_t))\}. \end{aligned} \tag{F.7}$$

The three terms account for price discretization, price selection, and inventory approximation, respectively. The first two are bounded by Eqs. (F.3) and (F.6), and Lemma 6 bounds the third by $C\varepsilon$. Combining these bounds gives

$$V_x - Q_x(p_t, \widetilde{y}_x(p_t)) \le C\sqrt{\frac{\ell_T(T^{-2})}{\max\{1, W_t(x,p_t)\}}} + C(\varepsilon + h^\alpha). \tag{F.8}$$

We sum Eq. (F.8) over the UCB rounds and apply Lemma 10 to the confidence-width term. Adding the profiling regret in Eq. (F.2) and the failure-event contribution gives

$$\text{Reg}_T \le C\,\text{polylog}(T)\left\{K_{\text{prof}}\delta_x^{-d}\varepsilon^{-1} + \sqrt{TK_{\text{dec}}h^{-d}} + T(\varepsilon + h^\alpha)\right\}. \tag{F.9}$$

For the Lipschitz price class, $K_{\mathrm{dec}} \le C\varepsilon^{-1}$, $K_{\mathrm{prof}} \le C\varepsilon^{-1/2}$, and $\delta_x^{-d} = h^{-d/2}$. With $\varepsilon = h^\alpha$, Eq. (F.9) becomes

$$\mathsf{Reg}_T \le C\,\mathrm{polylog}(T)\left\{h^{-(d+3\alpha)/2} + \sqrt{Th^{-d-\alpha}} + Th^\alpha\right\}.$$

Ignoring logarithmic factors, the choice $h = T^{-1/(d+3\alpha)}$ is equivalent to $T = h^{-(d+3\alpha)}$, and consequently

$$h^{-(d+3\alpha)/2} = \sqrt{T}, \qquad \sqrt{Th^{-d-\alpha}} = Th^\alpha = T^{(d+2\alpha)/(d+3\alpha)}.$$

Since $(d+2\alpha)/(d+3\alpha) > 1/2$, the profiling term is lower order. The logarithmic modification in Algorithm 1 changes only the polylogarithmic factor, completing the proof. □

Complete proof of Corollary 1. Let $r_{\mathrm{sm}} = (2d+3\alpha)/(2d+5\alpha) < 1$. With the smooth bandwidth and grid, $h_T \to 0$ and the profiling-action bound below is $o(T)$, so there is a finite $T_{0,\mathrm{sm}}$ after which the deterministic visibility fallback is inactive. Earlier horizons satisfy $R_{\max}T \le R_{\max}T_{0,\mathrm{sm}}^{1-r_{\mathrm{sm}}}T^{r_{\mathrm{sm}}}$. For later horizons, we apply the event bounds, profiling budget, optimism inequalities, and counting argument in Eqs. (F.1) and (F.9) and the intervening steps. Under the bounded-second-derivative condition, Proposition 4 permits a mesh of width $\sqrt{\varepsilon}$. It has $K_{\mathrm{dec}} = O(\varepsilon^{-1/2})$ and price approximation loss $O(\varepsilon)$. Therefore Eq. (F.9) becomes

$$\mathsf{Reg}_T \le C\,\mathrm{polylog}(T)\left\{h^{-d/2}\varepsilon^{-3/2} + \sqrt{Th^{-d}\varepsilon^{-1/2}} + T(\varepsilon + h^\alpha)\right\}.$$

Substitute $\varepsilon = h^\alpha$. Ignoring logarithms, the bandwidth $h = T^{-2/(2d+5\alpha)}$ satisfies $T = h^{-(d+5\alpha/2)}$, so

$$h^{-d/2}\varepsilon^{-3/2} = h^{-(d+3\alpha)/2},$$
$$\sqrt{Th^{-d}\varepsilon^{-1/2}} = h^{-d-3\alpha/2},$$
$$T\varepsilon = h^{-d-3\alpha/2}.$$

The profiling exponent $(d+3\alpha)/2$ is no larger than $d+3\alpha/2$, and is strictly smaller for $d > 0$. Finally,

$$h^{-d-3\alpha/2} = T^{(2d+3\alpha)/(2d+5\alpha)}.$$

The logarithmic factor in the stated bandwidth is again absorbed by $\mathrm{polylog}(T)$. □

G Proofs for the Lower Bound Construction

Lemma 11 (Slack flat-value baseline)**.** *There exist a smooth compactly supported mean-zero density f_0, intervals $\mathcal{P}_0 \Subset \mathrm{int}(\mathcal{P}_1) \Subset \mathcal{P}$, a smooth λ_0, and $C_0, s_0 > 0$ such that the constant family $F_x \equiv F_0$ has the residual-law properties of Assumption 1 and the density bounds of Assumption 3,*

$$I_H := 4\int_{\mathbb{R}} \left|\frac{d}{dz}\sqrt{f_0(z)}\right|^2 dz < \infty,$$

$G_0(p) = p\lambda_0(p) + C_{F_0}(p)$ equals C_0 on $\mathcal{P}_0$, is smaller outside $\mathcal{P}_0$, and is at most $C_0 - 4s_0$ outside $\mathcal{P}_1$. The baseline can be chosen so that, uniformly in p,

$$\lambda_0(p) + \underline{z}, \quad q_{0,\rho}(p) - (y_{\min} + r_y), \quad (y_{\max} - r_y) - y_0^*(p), \quad z_p^0 - z_\rho^0 - m_q \ge 4s_0,$$

with $0 < r_y \le r_q$, and G_0, λ_0 have uniformly bounded first two price derivatives.

Definition 3 (Normalized context–price packing)**.** *Fix the baseline and the interval $\mathcal{P}_0$ from Lemma 11. Let*

$$\zeta(t) = e^{-1/t}\mathbf{1}\{t > 0\}, \qquad \vartheta(u) = \frac{\zeta(1/2 - |u|)}{\zeta(1/2 - |u|) + \zeta(|u| - 1/4)}. \tag{G.1}$$

Then $\vartheta \in C^\infty(\mathbb{R}; [0,1])$, it equals one on $[-1/4, 1/4]$, and its support is $[-1/2, 1/2]$. Define the normalized context bump

$$\Psi(v) = \prod_{\ell=1}^{d} \vartheta(v_\ell), \qquad v \in \mathbb{R}^d.$$

For $0 < h \le 1/4$, put $m_h = \lfloor (2h)^{-1} \rfloor$, index the multi-indices $r \in \{0, \dots, m_h - 1\}^d$ by $j = 1, \dots, M_h$, where $M_h = m_h^d$, and set

$$\begin{aligned} c_{j,\ell} &= (2r_\ell + 1)h, \qquad \ell = 1, \dots, d, \\ B_j &= c_j + [-h/2, h/2]^d, \\ B_j^\circ &= c_j + [-h/4, h/4]^d, \\ \psi_j(x) &= \Psi\left(\frac{x - c_j}{h}\right). \end{aligned} \tag{G.2}$$

Thus the cells $B_j \subset [0,1]^d$ are pairwise disjoint, $\operatorname{supp}(\psi_j) = B_j$, and $\psi_j = 1$ on B_j°.

Choose once and for all a closed interval $J = [p_L, p_R] \Subset \operatorname{int}(\mathcal{P}_0)$, and write $L_0 = p_R - p_L$. For $0 < w \le L_0/4$, let

$$K_\Delta = \left\lfloor \frac{L_0}{2w} \right\rfloor, \qquad p_k = p_L + (2k-1)w, \qquad k = 1, \dots, K_\Delta, \tag{G.3}$$

and define

$$\varphi_0 \equiv 0, \qquad \varphi_k(p) = \vartheta\left(\frac{p - p_k}{w}\right), \qquad \mathcal{S}_k = \{p : \varphi_k(p) > 0\}. \tag{G.4}$$

The sets $\mathcal{S}_k$ are pairwise disjoint and contained in J. Finally, let

$$\Omega_0 = \{0, 1, \dots, K_\Delta\}^{M_h}, \qquad \Omega = \{1, \dots, K_\Delta\}^{M_h},$$

and, for every label vector $\omega \in \Omega_0$, define

$$g_\omega(x, p) = \Delta \sum_{j=1}^{M_h} \psi_j(x) \varphi_{\omega_j}(p). \tag{G.5}$$

The augmented set Ω_0 contains the zero-label reference experiments. The final hard prior is supported on Ω.

Proposition 6 (Properties of the normalized packing)**.** *For the construction in Definition 3, constants depending only on d, L_0, and the fixed bump ϑ satisfy*

$$c h^{-d} \le M_h \le C h^{-d}, \qquad c w^{-1} \le K_\Delta \le C w^{-1}. \tag{G.6}$$

Under the uniform context distribution,

$$\mathbb{P}(x_t \in B_j) = h^d, \qquad \mathbb{P}(x_t \in B_j^\circ) = 2^{-d} h^d. \tag{G.7}$$

Moreover, uniformly over $\omega \in \Omega_0$,

$$\begin{aligned} \|g_\omega\|_\infty &\le \Delta, \\ \|\partial_p g_\omega\|_\infty &\le C\Delta/w, \\ \|\partial_{pp} g_\omega\|_\infty &\le C\Delta/w^2, \\ |g_\omega(x,p) - g_\omega(x',p)| &\le C\Delta h^{-\alpha} \|x - x'\|^\alpha. \end{aligned} \tag{G.8}$$

For $x \in B_j^\circ$, the perturbation reduces exactly to

$$g_\omega(x, p) = \Delta \varphi_{\omega_j}(p). \tag{G.9}$$

Consequently, when $\omega_j = k \ge 1$, its maximum over $\mathcal{P}_0$ is Δ, attained on the plateau of φ_k, while it is zero for every $p \notin \mathcal{S}_k$.

Lemma 12 (Stability of lower-bound perturbations)**.** *There are $h_0, \Delta_0, c_\Delta > 0$ and a baseline from Lemma 11 such that, for $h \le h_0$, $\Delta = c_\Delta h^\alpha \le \Delta_0$, the packing in Definition 3 with $w = \Delta$, and every $\omega \in \Omega_0$,*

$$\lambda_\omega(x, p) = \lambda_0(p) + g_\omega(x, p)/p$$

is nonnegative, satisfies all assumptions, and has an optimal price in $\operatorname{int}(\mathcal{P}_0)$. The twice-smooth construction satisfies the same claim for the packing with $w = \sqrt{\Delta}$ and uniformly bounded $\partial_{pp} G_{\omega,x}$.

Proof of Lemma 11. Let

$$\widetilde{f}(u) = c_f \exp\left(-\frac{1}{1-u^2}\right) \mathbf{1}\{|u| < 1\}.$$

This density is symmetric and hence mean zero, is positive on every compact subset of $(-1, 1)$, and vanishes to all orders at the endpoints. Therefore $\sqrt{\widetilde{f}} \in H^1(\mathbb{R})$. A fixed rescaling $f_0(u) = a^{-1}\widetilde{f}(u/a)$ retains these properties and has finite I_H. Since $\rho < \inf_{p\in\mathcal{P}} \phi(p)$ and ϕ is continuous, after taking the price interval sufficiently small around an interior $p_0 > 0$, all relevant probability levels lie in one compact subset of $(0, 1)$. Their quantiles therefore lie in a compact subinterval on which f_0 admits uniform positive lower and finite upper bounds. Choose r_q, κ from this compact subinterval and a global density upper bound K. Write $z_u^0 = F_0^{-1}(u)$, $z_p^0 = z_{\phi(p)}^0$, and let C_{F_0} denote the residual-value functional in Proposition 3 for F_0. Choose $0 < m_q < (\inf_p \phi(p) - \rho)/K$, as in Lemma 3.

Choose a location level $\bar{\lambda}$, and fixed inventory endpoints, so that $\bar{\lambda} + \underline{z} > 0$ and both $\bar{\lambda} + z_\rho^0$ and $\bar{\lambda} + z_p^0$ lie strictly inside the inventory interval. This is possible by first taking the noise scale a small and then centering the inventory interval at $\bar{\lambda}$. Define

$$C_0 = p_0\bar{\lambda} + C_{F_0}(p_0), \qquad \lambda_{\text{flat}}(p) = \frac{C_0 - C_{F_0}(p)}{p}. \tag{G.10}$$

The function C_{F_0} is C^2 on the chosen price interval: the quantile map $p \mapsto z_p^0$ is C^2 there by the density lower bound, and the truncated expectations defining C_{F_0} may be differentiated under the integral. Hence λ_{flat} is C^2, equals $\bar{\lambda}$ at p_0, and remains inside all strict feasibility margins after shrinking the price interval.

Choose $\mathcal{P}_0 \Subset \text{int}(\mathcal{P}_1) \Subset \mathcal{P}$ around p_0 and a smooth cutoff $\chi : \mathcal{P} \to [0, 1]$ that is zero on $\mathcal{P}_0$ and one outside $\mathcal{P}_1$, and is strictly positive on $\mathcal{P} \setminus \mathcal{P}_0$. For a sufficiently small fixed $s_0 > 0$, set

$$G_0(p) = C_0 - 4s_0\chi(p), \qquad \lambda_0(p) = \frac{G_0(p) - C_{F_0}(p)}{p}. \tag{G.11}$$

Continuity on the compact price interval and the strict margins of λ_{flat} permit choosing s_0 and the interval widths so that all four inequalities in part (iii) hold. The cutoff construction gives part (ii), while $p \geq p_{\min} > 0$ and smoothness of G_0, C_{F_0} give the derivative bounds in part (iv). This proves every item of the lemma. □

Proof of Proposition 6. For $h \leq 1/4$, $(4h)^{-1} \leq m_h \leq (2h)^{-1}$, which yields the first cardinality bound. Similarly, $w \leq L_0/4$ gives $L_0/(4w) \leq K_\Delta \leq L_0/(2w)$. The cells and their cores have Lebesgue volumes h^d and $(h/2)^d$, respectively, proving Eq. (G.7) under the uniform context law.

The centers of distinct context cells differ by at least $2h$ in one coordinate, whereas every B_j has side length h. Hence their supports are disjoint. Likewise, adjacent price centers are $2w$ apart and each price support has width w, so the sets $\mathcal{S}_k$ are pairwise disjoint. The first three bounds in Eq. (G.8) now follow from $0 \leq \Psi, \vartheta \leq 1$, the disjoint context supports, and the fixed finite quantities $\|\vartheta'\|_\infty$ and $\|\vartheta''\|_\infty$.

It remains to verify the context Hölder bound, including pairs of points in different cells. Since Ψ is smooth and compactly supported,

$$|\psi_j(x) - \psi_j(x')| \leq C \min\{1, h^{-1}\|x - x'\|\} \leq Ch^{-\alpha}\|x - x'\|^\alpha. \tag{G.12}$$

When x, x' lie in the same cell, or one of them lies outside every cell, this proves the claim directly. When they lie in distinct cells, the separation of the supports gives $\|x - x'\| \geq h$. Boundedness by one then gives the same inequality. Multiplication by Δ proves the last bound in Eq. (G.8). Finally, $\psi_j = 1$ on B_j°, all other context bumps vanish there, and $\max_p \varphi_k(p) = 1$. This proves Eq. (G.9) and the final assertion. □

Proof of Lemma 12. We verify Assumptions 1 to 3 uniformly over the packing, and Assumption 4 for the smooth-price construction.

Context arrivals. The context distribution is uniform on $[0,1]^d$, so the i.i.d. requirement in Assumption 1 holds.

Common perturbation bound and slack choice. By Proposition 6, $|g_\omega(x,p)| \le \Delta$, and, with $\delta\lambda_\omega = g_\omega/p$,

$$\|\delta\lambda_\omega\|_\infty \le A_0\Delta, \qquad A_0 := p_{\min}^{-1}. \tag{G.13}$$

Fix

$$\Delta_0 \le \min\left\{\frac{s_0}{A_0},\ s_0,\ 1\right\}, \tag{G.14}$$

Choose $c_\Delta > 0$ small enough that the context Hölder constant below is at most the prescribed L_x. Then choose $h_0 \le 1/4$ so that, for every $h \le h_0$,

$$\Delta = c_\Delta h^\alpha \le \Delta_0, \qquad \sqrt{\Delta} \le L_0/4. \tag{G.15}$$

This also guarantees $w \le L_0/4$ in both price-regularity regimes, so every object in Definition 3 is well defined.

Location surface and nonnegative demand. The baseline margin and Eqs. (G.13) and (G.14) give

$$\lambda_\omega(x,p) + \underline{z} \ge 4s_0 - A_0\Delta \ge 3s_0 > 0.$$

The same inequality and compactness give uniform boundedness of demand.

The context bound in Eq. (G.8), together with $p \ge p_{\min}$ and $\Delta = c_\Delta h^\alpha$, gives

$$|\lambda_\omega(x,p) - \lambda_\omega(x',p)| \le C c_\Delta \|x - x'\|^\alpha.$$

For price regularity of the demand surface,

$$\partial_p\left(\frac{g_\omega(x,p)}{p}\right) = \frac{\partial_p g_\omega(x,p)}{p} - \frac{g_\omega(x,p)}{p^2}.$$

By Eq. (G.8), the Lipschitz packing has $|\partial_p g_\omega| \le C\Delta/w \le C$ because $w = \Delta$. In the smooth packing, $w = \sqrt{\Delta}$, and the same first derivative is even $O(\sqrt{\Delta})$. Combining these bounds with the fixed smooth baseline proves the price-Lipschitz part of Assumption 2 uniformly over the two families.

Residual distribution and quantile regularity. Set $F_x \equiv F_0$ for every context and every packing instance. This constant family has Hölder modulus zero. Its mean-zero normalization, compact support, bounds K, κ, r_q, and quantiles z_ρ^0, z_p^0 are identical across contexts and across the packing. Thus the residual-law and density conditions in Assumptions 1 and 3 hold, and the noise part of Assumption 2 holds with $L_F = 0$. Shocks are drawn freshly from F_0 after each action. The additional slack in Lemma 11 gives

$$z_p^0 - z_\rho^0 \ge m_q + 4s_0 \ge m_q$$

uniformly in p, which will be used for inventory visibility below.

Inventory feasibility. The stockout probability is

$$S_{\omega,x}(y \mid p) = 1 - F_0(y - \lambda_\omega(x,p)).$$

Its root is

$$y^*_{\omega,x}(p) = \lambda_\omega(x,p) + z_p^0,$$

and the calibration quantile is

$$q_{\rho,\omega}(x,p) = \lambda_\omega(x,p) + z_\rho^0.$$

Both quantities shift from their baseline values by the same $\delta\lambda_\omega(x,p)$. Hence the explicit slack choice gives

$$\begin{aligned} q_{\rho,\omega}(x,p) - (y_{\min} + r_y) &\ge 4s_0 - A_0\Delta \ge 3s_0, \\ (y_{\max} - r_y) - y^*_{\omega,x}(p) &\ge 4s_0 - A_0\Delta \ge 3s_0, \\ y^*_{\omega,x}(p) - q_{\rho,\omega}(x,p) &= z_p^0 - z_\rho^0 \ge m_q + 4s_0. \end{aligned} \tag{G.16}$$

These inequalities prove Assumption 3. The associated uniqueness, local slope, and quadratic inventory-loss properties then follow from Lemmas 1 and 2.

Optimal-price interiority. By Proposition 3, for every price,

$$G_{\omega,x}(p) = p\lambda_\omega(x,p) + C_{F_0}(p) = G_0(p) + g_\omega(x,p).$$

All price bumps are supported in $J \Subset \mathrm{int}(\mathcal{P}_0)$. For a context with positive bump amplitude, a point on the corresponding price-bump plateau attains $C_0 + \Delta\psi_j(x) > C_0$, whereas every price outside $\mathcal{P}_0$ has value at most C_0. With zero bump amplitude, every point of $\mathcal{P}_0$, hence some interior point, is a baseline maximizer. Thus every perturbed instance has an optimal price in the required interior interval, and no price-monotonicity property is used.

Price-value regularity. In the Lipschitz construction, Eq. (G.8) gives $|\partial_p g_\omega| \le C\Delta/w = O(1)$, yielding the uniform Lipschitz value bound after adding the fixed smooth baseline G_0. In the smooth construction, $w = \sqrt{\Delta}$, so

$$|\partial_{pp} g_\omega(x,p)| \le C\Delta/w^2 = O(1),$$

which proves the bounded-second-derivative specialization. Outside $\mathcal{P}_0$, the perturbation vanishes and G_0 has fixed bounded derivatives. This completes the uniform verification of Assumptions 1 to 3, together with Assumption 4 for the smooth-price packing. □

Lemma 13 (Location-family information bound)**.** *With $H^2(P,Q) = \frac{1}{2}\int(\sqrt{dP} - \sqrt{dQ})^2$, shifts of the baseline density satisfy*

$$H^2(f_0(\cdot - u), f_0(\cdot - v)) \le \frac{I_H}{4}|u - v|^2.$$

Thus neighboring bump instances have one-round squared Hellinger distance $O(\Delta^2)$ on the affected context–price region and zero elsewhere. The same bound holds for censored sales under every adaptive inventory action.

Proof. Let $s = \sqrt{f_0}$. The fundamental theorem of calculus gives

$$H^2(f_0(\cdot - u), f_0(\cdot - v)) \le \frac{1}{2}|u - v|^2 \|s'\|_2^2 = \frac{I_H}{8}|u - v|^2.$$

For the neighbors used below, ω and ω' agree outside one cell j, while their labels in that cell are 0 and k. By Eq. (G.5), their locations agree unless $x \in B_j$ and $p \in \mathcal{S}_k$, and everywhere differ by at most $|g_\omega(x,p) - g_{\omega'}(x,p)|/p_{\min} \le C\Delta$. Censoring is a measurable image of demand, so data processing proves the claim. □

H Complete Direct-Sum Proof of the Lower Bound

We use the following sequential form of squared Hellinger subadditivity. Let $P_t(\cdot \mid z_{<t})$ and $Q_t(\cdot \mid z_{<t})$ be probability kernels on standard Borel spaces, and let P, Q be their induced laws on $Z_1, \ldots, Z_n$. With the normalization in Lemma 13,

$$H^2(P,Q) \le 7\mathbb{E}_P\left[\sum_{t=1}^{n} H^2(P_t(\cdot \mid Z_{<t}), Q_t(\cdot \mid Z_{<t}))\right]. \tag{H.1}$$

This is Lemma F.2 of Foster et al. (2024). The inequality permits adaptive kernels and an expectation under just one reference law. Its absolute factor is absorbed into the constants below.

Complete proof of Theorem 4. Fix one of the two price-regularity regimes, set $w = \Delta$ in the Lipschitz regime and $w = \sqrt{\Delta}$ in the twice-smooth regime, and use the family Ω in Definition 3. By Lemma 12, every member of this family belongs to the claimed model class. It remains to show that no adaptive policy can learn all context-cell labels quickly enough. We give the argument for a deterministic policy. Conditioning on a randomized policy's internal seed and then applying Yao's principle gives the stated conclusion for randomized policies.

All objects below are those formally constructed in Definition 3. In particular, the price supports $\mathcal{S}_k$ are pairwise disjoint and $\psi_j = 1$ on B_j°. We use the three counts

$$\begin{aligned} \overline{N}_{j,k} &= \sum_{t=1}^{T} \mathbf{1}\{x_t \in B_j,\ p_t \in \mathcal{S}_k\}, \\ N_j^\circ &= \sum_{t=1}^{T} \mathbf{1}\{x_t \in B_j^\circ\}, \\ N_{j,k}^\circ &= \sum_{t=1}^{T} \mathbf{1}\{x_t \in B_j^\circ,\ p_t \in \mathcal{S}_k\}. \end{aligned} \tag{H.2}$$

The first count includes the transition shell $B_j \setminus B_j^\circ$. This inclusion is essential because observations in that shell also carry information about the label. The core counts will be used only to lower-bound regret.

Fix a cell j. Put an independent uniform prior on the labels of all other cells. Let $\mathbb{P}^{(j,0)}$ be the resulting mixture history law when cell j has no bump, and let $\mathbb{P}^{(j,k)}$ be the mixture law when its label is k. The no-bump instance serves as the reference experiment. The final prior is supported on the positive-label instances. Conditional on a fixed configuration of the other labels, Eq. (H.1) and Lemma 13 give

$$H^2(\mathbb{P}^{(j,0)}_{\omega_{-j}}, \mathbb{P}^{(j,k)}_{\omega_{-j}}) \le C\Delta^2 \mathbb{E}^{(j,0)}_{\omega_{-j}}[\overline{N}_{j,k}]. \tag{H.3}$$

Apply this inequality to the sequential variables

$$(x_1, (p_1, y_1), O_1, \ldots, x_T, (p_T, y_T), O_T).$$

The context kernels are the same uniform law in both experiments. Conditional on a common history and the current context, the policy also uses the same action kernel. Their Hellinger contributions in Eq. (H.1) are therefore zero. For the observation step, conditional on a common history and action, the two locations differ by $\Delta\psi_j(x_t)\varphi_k(p_t)/p_t$. Thus Lemma 13 bounds that step's squared Hellinger distance by

$$C\Delta^2 \mathbf{1}\{x_t \in B_j,\ p_t \in \mathcal{S}_k\}.$$

This bound includes every feasible inventory: the same censoring map $D \mapsto \min\{D, y_t\}$ is applied in the two conditional experiments. Taking P in Eq. (H.1) to be the zero-label reference law, summing, and absorbing the factor 7 proves Eq. (H.3). The history spaces here are finite products of Euclidean spaces, so the required conditional kernels exist.

Squared Hellinger distance is jointly convex. Averaging Eq. (H.3) over the other labels and then over k gives

$$\begin{aligned} \frac{1}{K_\Delta} \sum_{k=1}^{K_\Delta} H^2(\mathbb{P}^{(j,0)}, \mathbb{P}^{(j,k)}) &\le \frac{C\Delta^2}{K_\Delta} \mathbb{E}^{(j,0)} \left[\sum_{k=1}^{K_\Delta} \overline{N}_{j,k} \right] \\ &\le C \frac{Th^d\Delta^2}{K_\Delta}. \end{aligned} \tag{H.4}$$

The second inequality uses disjointness of the price supports and $\mathbb{P}(x_t \in B_j) = h^d$, as established in Proposition 6. Choose the fixed packing constant c_Δ and the bandwidth so that

$$\frac{Th^d\Delta^2}{K_\Delta} \le c_{\inf} \tag{H.5}$$

for a sufficiently small numerical $c_{\inf}$. Since $\mathrm{TV}(P, Q) \le \sqrt{2H^2(P, Q)}$, Jensen's inequality and Eq. (H.4) yield

$$\frac{1}{K_\Delta} \sum_{k=1}^{K_\Delta} \mathrm{TV}(\mathbb{P}^{(j,0)}, \mathbb{P}^{(j,k)}) \le c_{\mathrm{tv}}, \tag{H.6}$$

where c_{tv} can be made smaller than $1/8$.

We next convert indistinguishability into regret. Define

$$A_{j,k} = \{N_j^\circ > 0,\ N_{j,k}^\circ \ge N_j^\circ/2\}, \qquad E_{j,k} = A_{j,k}^c \cap \{N_j^\circ \ge c_B Th^d\}.$$

Because the price supports are disjoint, on every sample path with $N_j^\circ > 0$, at most two of the events $A_{j,k}$ can occur. Therefore

$$\frac{1}{K_\Delta}\sum_{k=1}^{K_\Delta}\mathbb{P}^{(j,0)}(A_{j,k}) \le \frac{2}{K_\Delta}. \tag{H.7}$$

By Eq. (G.7), the random variable N_j° is binomial with mean $2^{-d}Th^d$, independently of the policy. For the bandwidths chosen at the end of the proof, this mean diverges: $Th^d \asymp T^{3\alpha/(d+3\alpha)}$ in the Lipschitz regime and $Th^d \asymp T^{5\alpha/(2d+5\alpha)}$ in the smooth regime. Taking $c_B = 2^{-d-1}$, the binomial lower-tail bound gives

$$\mathbb{P}^{(j,0)}(N_j^\circ < c_B Th^d) \le \exp\{-2^{-d}Th^d/8\}.$$

In particular, for all sufficiently large T,

$$\mathbb{P}^{(j,0)}(N_j^\circ \ge c_B Th^d) \ge \frac{3}{4}. \tag{H.8}$$

Equations Eqs. (H.6) to (H.8) imply

$$\frac{1}{K_\Delta}\sum_{k=1}^{K_\Delta}\mathbb{P}^{(j,k)}(E_{j,k}) \ge \frac{3}{4} - \frac{2}{K_\Delta} - c_{\mathrm{tv}} \ge \frac{3}{8} =: c_E. \tag{H.9}$$

Indeed, the average reference probability is at least $3/4 - 2/K_\Delta$, and changing from the reference to the alternatives reduces it by at most the average total-variation distance. Here we take T large enough that $K_\Delta \ge 8$ and use $c_{\mathrm{tv}} \le 1/8$.

By Eq. (G.9), on B_j° under label $\omega_j = k$ and for $p \in \mathcal{P}_0$, the optimized value is

$$G_{\omega,x}(p) = C_0 + \Delta\varphi_k(p).$$

Its maximum is $C_0 + \Delta$. Every price in $\mathcal{P}_0$ but outside $\mathcal{S}_k$ has gap exactly Δ. Every price outside $\mathcal{P}_0$ has gap at least Δ, because the baseline extension is at most C_0 and all price bumps are supported in $\mathcal{P}_0$. On $E_{j,k}$, at least $N_j^\circ/2$ core visits use prices outside $\mathcal{S}_k$. Inventory suboptimality is nonnegative, so

$$\begin{aligned}\mathsf{Reg}_j &:= \sum_{t:x_t\in B_j^\circ}\{V_{x_t} - Q_{x_t}(p_t, y_t)\}\\ &\ge \frac{\Delta}{2}N_j^\circ \mathbf{1}\{E_{j,k}\}.\end{aligned}$$

Combining this inequality with Eq. (H.9) proves the cellwise Bayes bound

$$\frac{1}{K_\Delta}\sum_{k=1}^{K_\Delta}\mathbb{E}^{(j,k)}[\mathsf{Reg}_j] \ge cTh^d\Delta. \tag{H.10}$$

Now draw ω uniformly from $\Omega = \{1,\dots,K_\Delta\}^{M_h}$. Equivalently, all cell labels are independent and uniform. For fixed j, integrating out the other labels gives exactly the mixture experiment above. The conditional observation kernel can depend directly on ω_j only on rounds with $x_t \in B_j$, which is precisely the localization used in Eq. (H.3). The core cells are disjoint, so summing Eq. (H.10) yields Bayes regret

$$\mathbb{E}_{\omega\sim\mathrm{Unif}(\Omega)}[\mathsf{Reg}_T] \ge cM_hTh^d\Delta \ge cT\Delta. \tag{H.11}$$

Yao's principle turns Eq. (H.11) into a worst-instance lower bound for every randomized policy.

It remains to solve the information balance. In the Lipschitz construction, $w = \Delta$ and $K_\Delta \asymp \Delta^{-1}$, so Eq. (H.5) is tight at

$$Th^d\Delta^3 \asymp 1.$$

Together with $\Delta = c_\Delta h^\alpha$, this gives

$$h \asymp T^{-1/(d+3\alpha)}, \qquad T\Delta \asymp T^{(d+2\alpha)/(d+3\alpha)}.$$

In the twice-smooth construction, $w = \sqrt{\Delta}$ and $K_\Delta \asymp \Delta^{-1/2}$. Hence

$$Th^d\Delta^{5/2} \asymp 1, \qquad h \asymp T^{-2/(2d+5\alpha)},$$

and

$$T\Delta \asymp T^{(2d+3\alpha)/(2d+5\alpha)}.$$

This proves both lower bounds for all $T \geq T_0$, with a fixed T_0. To cover the remaining horizons, fix one admissible packing with $K_\Delta \geq 2$ and positive, fixed h, Δ. Before the first observation, the independent uniform label of a core cell is unknown to the policy. Conditional on any first-round context in that core, the probability of selecting its correct price support, averaged over labels, is at most $1/K_\Delta$. Consequently the first-round Bayes regret is at least

$$c_* := \Delta(1 - 1/K_\Delta)M_h 2^{-d} h^d > 0,$$

uniformly over policies. Subsequent regret is nonnegative. For either claimed exponent r, reducing the lower-bound constant to at most c_*/T_0^r therefore covers every integer $1 \leq T < T_0$. □

I Experimental Details

I.1 Additional Details for the Main Experiment

Demand and noise environment. We generate latent demand from the additive location model

$$D_t = \lambda_{\alpha_0}(x_t, p_t) + \epsilon_t, \qquad \epsilon_t \sim \mathrm{Unif}[-a, a].$$

To construct an α_0-Hölder-smooth demand surface, define the centered cusp

$$\psi_{c,\alpha_0}(u) = |u - c|^{\alpha_0} - \frac{c^{\alpha_0+1} + (1-c)^{\alpha_0+1}}{\alpha_0 + 1}.$$

For $d = 1$, the location surface is

$$\lambda_{\alpha_0}(x, p) = 1.55 - 0.58\psi_{0.35,\alpha_0}(x) + 0.22\psi_{0.78,\alpha_0}(x) - 0.42p + 0.16p\psi_{0.62,\alpha_0}(x).$$

For $d = 2$, it is

$$\begin{aligned}\lambda_{\alpha_0}(x, p) = 1.55 &- 0.42\psi_{0.32,\alpha_0}(x_1) + 0.28\psi_{0.79,\alpha_0}(x_1) \\ &- 0.36\psi_{0.41,\alpha_0}(x_2) + 0.24\psi_{0.73,\alpha_0}(x_2) - 0.42p \\ &+ 0.10p\psi_{0.61,\alpha_0}(x_1) - 0.08p\psi_{0.27,\alpha_0}(x_2).\end{aligned}$$

Each surface is α_0-Hölder and contains an α_0-order cusp. The chosen action produces censored sales $O_t = \min\{D_t, y_t\}$. The learner observes only O_t.

Parameter choices and data generation. The action domains are $p \in [0.1, 1]$ and $y \in [0, 2.2]$. We set the shortage cost to $b = 0.4$, the holding cost to $h_c = 1$, the visible quantile level to $\rho = 0.2$, and the noise half-width to $a = 0.5$. The scaling experiment uses $T \in \{4000, 8000, 16000, 32000, 64000\}$ and five repetitions for each configuration $(d, \alpha_0) \in \{(1, 0.5), (1, 0.75), (1, 1), (2, 0.5), (2, 0.75), (2, 1)\}$.

For each repetition, contexts are drawn i.i.d. uniformly from $[0, 1]^d$ and residual shocks are drawn independently from $\mathrm{Unif}[-0.5, 0.5]$. The complete context and residual-shock arrays are generated before the policy starts, using independent child random streams.

For replication r under configuration (d, α_0, T), the random seed is set to $20260824 + 7919d + 1000\alpha_0 + 9176T + 1009r$.

The algorithm is supplied with the true environmental smoothness, $\widehat{\alpha} = \alpha_0$. At declared horizon T, its bandwidth and profiling parameters are

$$h = \left(\frac{\log(eT)}{T}\right)^{1/(d+3\alpha)}, \qquad \varepsilon = h^\alpha, \qquad \delta_x = \varepsilon^{1/(2\alpha)}.$$

Here $\alpha = \widehat{\alpha} = \alpha_0$. The decision-price mesh is ε, the profiler-anchor mesh is $\sqrt{\varepsilon}$, and profiler cells have diameter at most δ_x. Each midpoint threshold uses a batch of

$$m_y = \left\lceil C_{\mathrm{tune}} \eta^{-2} \log(T|\Gamma_y||\mathcal{Z}_{\mathrm{prof}}|B_y/T^{-2}) \right\rceil$$

observations, where $\eta = c_\eta \sqrt{\varepsilon}$. We set $C_{\mathrm{tune}} = 0.05$, $c_\eta = 0.25$, and $h_{\mathrm{vis}} = 1$. C_{tune} is a fixed scale factor in batch lengths and confidence radii. The reported values are chosen empirically. The theoretical confidence guarantees require the tuning conditions in Appendix C.1. The value h_{vis} is only a technical startup cutoff. Every experimental bandwidth satisfies $h \leq 1$, so this cutoff never changes an action in the reported experiments.

Simulation protocol and evaluation. We run the Lipschitz-price mode of Algorithm 1. Within each unfinished profiler cell, anchors are visited round-robin. A midpoint inventory threshold is held fixed for m_y observations and compared with $\tau(p) \pm 2\eta$. A decision price uses the completed inventory estimate of its nearest anchor. Completed cells switch immediately to calibrated kernel UCB, whereas unfinished cells continue profiling on their own arrivals. Only UCB rounds enter the price-specific proxy and sales-quantile histories.

The regret benchmark maximizes $G_x(p)$ on an 801-point price grid over $[0.1, 1]$. One-round regret is the expected-value gap

$$G^*_{x_t} - Q_{x_t}(p_t, y_t),$$

. Reported curves are means across repetitions. Shaded bands are mean plus or minus 1.96 standard errors. The top row of Fig. 1 reports $d = 1$ and the bottom row reports $d = 2$. The dashed guide is proportional to $T^{\beta(d,\alpha)}$ and is normalized at the first plotted horizon.

I.2 Twice-Smooth Price Experiment

We evaluate the twice-smooth price specialization of Corollary 1 using the controlled demand environment, six (d, α_0) configurations, horizons, and oracle regret benchmark from Appendix I.1. We run five trajectories for each configuration and horizon, using base seed 2123345 and the same configuration-dependent seed offsets as in the main experiment. For the controlled uniform-noise model, the optimized price-value curve is twice continuously differentiable on the compact price interval, with a uniformly bounded second derivative, so this environment belongs to the specialized class in Assumption 4.

With $\alpha = \widehat{\alpha} = \alpha_0$, the twice-smooth mode sets

$$h = \left(\frac{\log(eT)}{T} \right)^{2/(2d+5\alpha)}, \qquad \varepsilon = h^\alpha, \qquad \Delta_p = \sqrt{\varepsilon}.$$

The profiler retains context resolution $\delta_x = \varepsilon^{1/(2\alpha)}$, price-anchor mesh $\sqrt{\varepsilon}$, and the same asynchronous profiling protocol. Hence the decision grid and profiler-anchor grid coincide in this mode. We use $C_{\mathrm{tune}} = 0.01$, $c_\eta = 0.3$, and $h_{\mathrm{vis}} = 1$. The first two constants differ from those in the main experiment, so the comparison reflects changes in both the price grid and tuning constants. The dashed guide in each panel is proportional to

$$T^{\beta_{\mathrm{sm}}(d,\alpha)}, \qquad \beta_{\mathrm{sm}}(d, \alpha) = \frac{2d + 3\alpha}{2d + 5\alpha}.$$

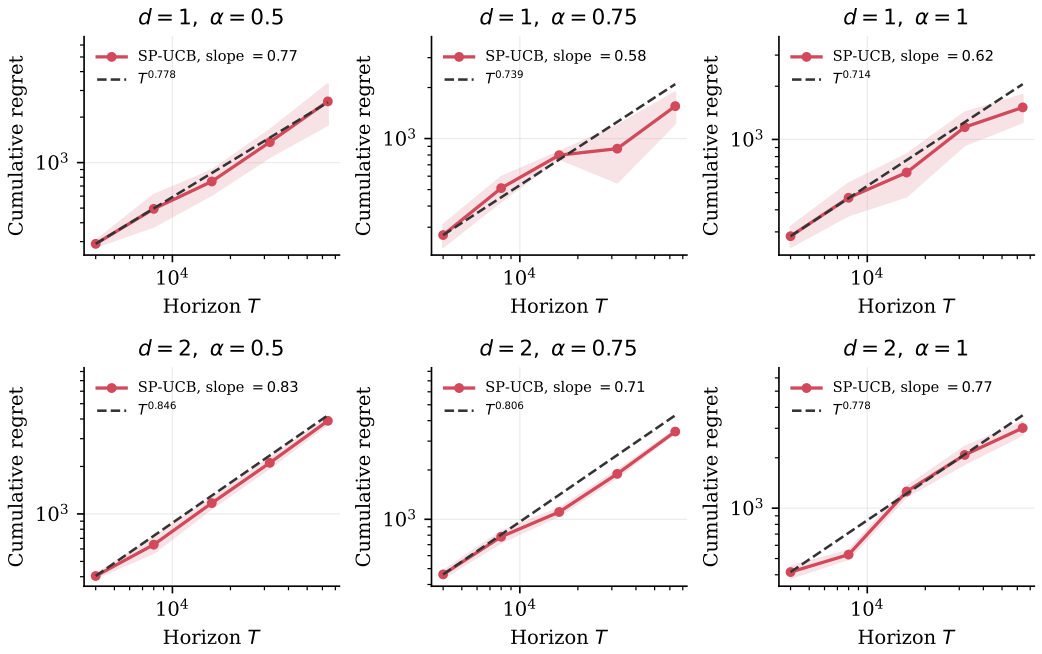


Figure 2: Cumulative-regret scaling in the twice-smooth price mode. Curves are means over five trajectories. Dashed curves are representative $T^{\beta_{\mathrm{sm}}}$ guides normalized at the first plotted horizon.

I.3 Comparisons with Neighboring Online Methods

We compare with two online joint pricing–inventory methods developed for a repeated, noncontextual demand environment: the CWZ trisection method of Chen et al. (2024) and a SALA-style implementation based on Chen et al. (2021). We construct context-agnostic baselines by pooling observations across contexts and applying CWZ and SALA to the resulting noncontextual problem. The contextual method of Han et al. (2026a) is discussed in Section 2.

Demand and residual environment. Contexts are drawn i.i.d. uniformly from $[0,1]^2$. Define

$$s(x) = 0.25x_1 + 0.25x_2 + 0.5x_1x_2, \qquad w(x) = 0.70 + 0.45s(x).$$

Demand follows

$$D_t = \lambda(x_t, p_t) + Z_t, \qquad \lambda(x,p) = 2.15 + \frac{1.70}{1 + 2\exp(16(p - w(x)))}, \qquad Z_t \sim \mathrm{Unif}[-0.15, 0.15].$$

The residuals are independent of context and price and have mean zero. The interaction term x_1x_2 shifts the location of the smooth demand response to price. Demand decreases strictly with price, while both the contextual response and the price response are nonlinear. Since $2.15 < \lambda(x,p) < 3.85$, realized demand lies in $[2,4]$ without clipping. The comparison evaluates context-pooled implementations on this contextual instance.

Parameter choices and data generation. All methods use $p \in [0.1, 1]$ and $y \in [0, 5]$. The shortage cost is $b = 0.4$, the holding cost is $h_c = 1$, the visible quantile level is $\rho = 0.2$, and $\alpha = 1$. We use $T = 16000$ and five independent repetitions. For this comparison, SP-UCB uses $C_{\mathrm{tune}} = 0.01$ and a profiling precision multiplier $c_\eta = 0.4$:

$$h = \varepsilon = 0.2317, \qquad \delta_x = \sqrt{\varepsilon} = 0.4813, \qquad \eta = c_\eta\sqrt{\varepsilon} = 0.1925.$$

The context partition and profiler anchor mesh retain their original resolution. There are nine profiling cells, five decision prices, and three profiler anchors. The stopping thresholds use $\tau(p) \pm 2\eta$, and the batch size is computed using the same η, giving ten observations per midpoint threshold.

The multiplier C_{tune} is used in the profiling batch size, the interval-width stopping condition, and the UCB bonus. These are empirical tuning choices.

The base seed is 20260827. Repetition $r \in \{0, \dots, 4\}$ uses seed $20260827 + 9176T + 1009r$. Independent child random streams generate the complete context array and uniform residual array before any policy starts. Within each repetition, all three methods receive the same (x_t, Z_t) on each calendar round.

Comparator implementations and evaluation. CWZ uses minimum epoch size 8, sample scale 0.006, and final exploitation fraction 0.35. The SALA-style implementation explores seven prices with exploration scale 1.1, estimates a pooled demand curve and residual law, and then exploits one price–inventory pair. Both comparators pool observations across contexts. The regret benchmark maximizes $G_x(p)$ on an 801-point price grid over $[0.1, 1]$. Curves report means over the five repetitions, with shaded bands showing ± 1.96 standard errors.

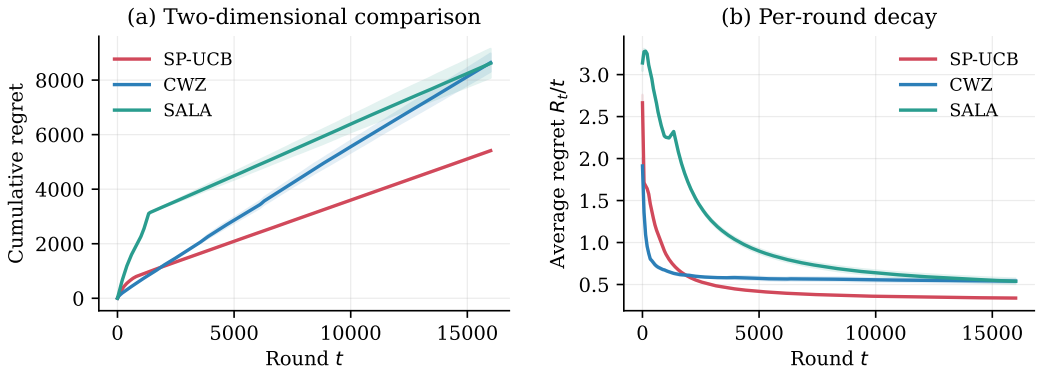


Figure 3: Comparison on the nonlinear interaction instance with $y \in [0, 5]$, $T = 16000$, $C_{\text{tune}} = 0.01$, and $c_\eta = 0.4$. Curves are means over five common-random-number repetitions.

In Fig. 3, SALA incurs substantial initial regret while exploring prices with high inventory to reconstruct demand, followed by approximately linear accumulation after committing to a fixed price–inventory pair. SP-UCB also incurs initial inventory-profiling costs, but subsequent contextual UCB learning yields a lower empirical regret slope. CWZ maintains a higher slope throughout much of the horizon: its search phase consumes 65% of the rounds, while its final policy retains a substantial per-round gap. This behavior partly reflects the current implementation, which defaults to midpoint inventory when the search budget is exhausted.

I.4 Smoothness-Mismatch Experiment

We use the same cusp demand family and simulation protocol as in Appendix I.1, but decouple the environmental smoothness from the smoothness supplied to the algorithm. The experiment crosses

$$(d, \alpha_0) \in \{1, 2\} \times \{0.5, 0.75, 1\}, \qquad \widehat{\alpha} \in \{0.5, 0.75, 1\},$$

over horizons $T \in \{4000, 8000, 16000, 32000\}$, with one repetition for each configuration. The environment uses the true cusp exponent α_0, whereas the algorithm constructs its bandwidth, price mesh, profile mesh, and confidence radii using $\widehat{\alpha}$. All other inputs are held fixed: we use the Lipschitz-price mode, $C_{\text{tune}} = 0.05$, $c_\eta = 0.3$, $h_{\text{vis}} = 1$, visible quantile level $\rho = 0.2$, shortage cost $b = 0.4$, holding cost $h_c = 1$, and noise half-width $a = 0.5$.

For fixed (d, α_0, T, r), all three values of $\widehat{\alpha}$ receive the same pre-generated context and residual-shock arrays. The complete context and residual-shock arrays are generated before the policy starts using independent child random streams, as in the main scaling experiment.

For reproducibility, the mismatch base seed is 20260903, with the same configuration-dependent offsets as in the main experiment, independently of $\widehat{\alpha}$. The regret benchmark is computed as in

Appendix I.1. The dashed reference curve in each panel is proportional to $T^{\beta(d,\alpha_0)}$, normalized at the first plotted horizon, where

$$\beta(d, \alpha_0) = \frac{d + 2\alpha_0}{d + 3\alpha_0},$$

the exponent obtained when the true smoothness governs the rate.

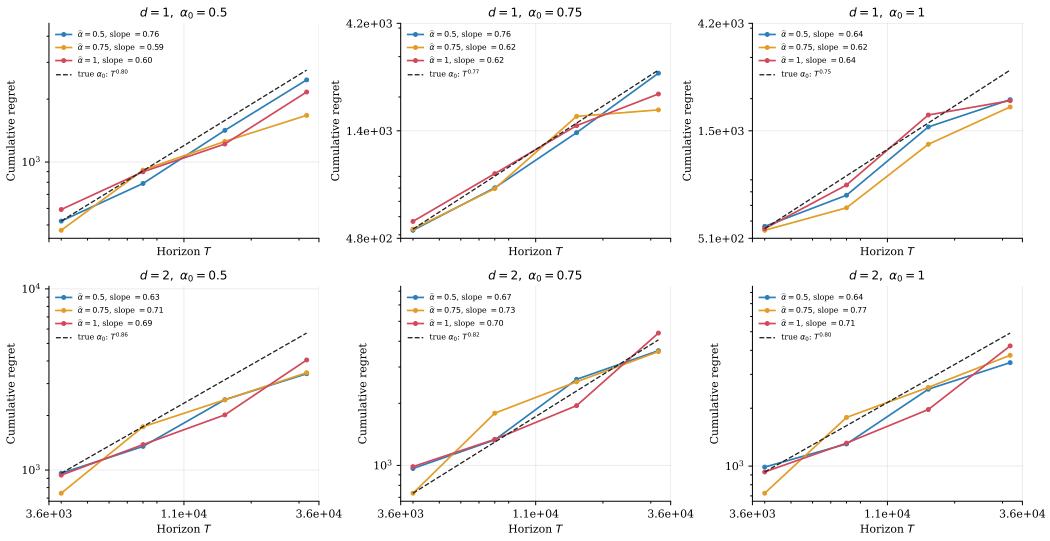


Figure 4: Scaling under smoothness-input mismatch. Each panel fixes the true cusp exponent α_0 and dimension d, and each colored curve uses a different algorithm input $\widehat{\alpha}$. The dashed black curve is a $T^{\beta(d,\alpha_0)}$ guide computed from the true smoothness and normalized to the matched-input curve at $T = 4000$. Each configuration uses one trajectory. The bands therefore have zero width.

At $T = 32000$, the input $\widehat{\alpha} = 0.75$ has the lowest regret in all three one-dimensional panels. In two dimensions, the lowest-regret inputs are 0.5, 0.75, and 0.5 for true smoothness 0.5, 0.75, and 1, respectively. $\widehat{\alpha} = 1$ has the highest regret in all three panels. The fitted log–log slopes range from 0.59 to 0.77 across the 18 curves. The lowest-regret smoothness input therefore varies with the configuration and horizon. These 72 runs describe finite-horizon behavior along one trajectory per configuration.

Changing $\widehat{\alpha}$ changes the bandwidth, decision-price grid, and profiling resolution simultaneously, so the observed curves reflect their combined effects. At the level of the theoretical model, an α_0-Hölder function is also $\widehat{\alpha}$-Hölder for $\widehat{\alpha} \leq \alpha_0$, with an adjusted constant on the compact domain. Conservative smoothness inputs therefore satisfy the smoothness requirement, with the remaining tuning conditions given in Appendix C.1. For $\widehat{\alpha} > \alpha_0$, the true smoothing bias may exceed the supplied bias correction and fall outside the scope of the current analysis.

J Inventory Error in Price Evaluation

Inventory-learning error affects the price evaluations in Chen et al. (2024) and our method differently. Their Algorithm 2 estimates profit differences between pairs of prices. Our method constructs a calibrated score for each price that can be reused across comparisons. In the setting studied below, inventory error introduces a first-order term into their pairwise correction. Our fixed low-quantile correction remains valid whenever the quantile is visible, so inventory error affects our score only through the quadratic profit loss near the optimal inventory. This allows coarse inventory estimates to support fine price comparisons.

J.1 Population Sensitivity

To isolate the effect of inventory error on price comparisons, fix a context x and suppress the context subscript. Under the location model, $D(p) = \lambda(p) + \xi$, where $\xi \sim F$, the fixed quantile satisfies $q_\rho(p) = \lambda(p) + z_\rho$. Since $\mathbb{E}[Y(p, y, O)] = Q(p, y) + b\lambda(p)$, our population score is

$$J(p, y) := \mathbb{E}[Y(p, y, O)] - bq_\rho(p) = Q(p, y) - bz_\rho. \tag{J.1}$$

For $y > q_\rho(p)$, the sales quantile equals the demand quantile, so the correction is identifiable and inventory error enters J only through the profit loss in Q.

For two fixed prices $p < p'$, let $b, h_c > 0$ and write

$$C_p := p + b + h_c, \qquad \tau(p) := h_c/C_p, \qquad z_p := F^{-1}(1 - \tau(p)), \qquad y^*(p) = \lambda(p) + z_p.$$

Take feasible inventories $y = y^*(p) + e$, $y' = y^*(p') + e'$, and define

$$\ell_p(e) := G(p) - Q(p, y^*(p) + e), \qquad L(v) := \mathbb{E}[(\xi - v)^+], \qquad \Delta G := G(p') - G(p).$$

Let $\Delta_{\mathrm{SP}} = J(p', y') - J(p, y)$, and let Δ_{C2} denote the population limit of CWZ2 at these same inventories.

We compare these population scores at the prescribed feasible inventories. The fixed-context instances also satisfy the decreasing, twice-differentiable demand and noise-density conditions of Chen et al. (2024).

Proposition 7 (Sensitivity to inventory-learning error)**.** *For the two-price comparison above, with visible low quantiles and $z_{p'} + e' > z_p$, the estimation errors satisfy, as $e, e' \to 0$,*

$$\Delta_{\mathrm{SP}} - \Delta G = O(e^2 + e'^2), \qquad \Delta_{\mathrm{C2}} - \Delta G = b\tau(p)e + O(e^2 + e'^2). \tag{J.2}$$

Proof. On the common fixed-context model, Assumptions 1 and 3 provide bounded density and interior critical inventories. The additional price and density regularity imposed by Chen et al. (2024) places this instance in their model class. The identities below follow from the shared-noise and inventory assumptions. Visibility identifies the low quantile used in J. Define $A_p(e) = \int_0^e [F(z_p + t) - F(z_p)]\, dt$. The density bound gives $0 \le A_p(e) \le \kappa' e^2/2$, for either sign of e. Integrating $\partial_y Q(p, y) = C_p[F(z_p) - F(y - \lambda(p))]$ yields $\ell_p(e) = C_p A_p(e)$. Together with Eq. (J.1), this gives

$$\Delta_{\mathrm{SP}} - \Delta G = \ell_p(e) - \ell_{p'}(e'). \tag{J.3}$$

For CWZ2, put $a = z_p + e$, $a' = z_{p'} + e'$. The high-price remaining inventory is $\delta' = (a' - \xi)^+$. CWZ2 inverts its distribution at $\tau(p)$, obtaining $u = a' - z_p > 0$:

$$\mathbb{P}(\delta' \le a' - z_p) = \mathbb{P}(\xi \ge z_p) = \tau(p).$$

The local density lower bound at z_p makes this inversion unique. Its correction is therefore

$$b\,[u\tau(p) - \mathbb{E}[\delta' \mathbf{1}\{0 < \delta' \le u\}]] = b[L(z_p) - L(a')].$$

The correction needed at the actual inventories is $b[L(a) - L(a')]$, since $Q(p, y) = \mathbb{E}[pO - h_c(y - O)] - bL(a)$. Their difference gives

$$\Delta_{\mathrm{C2}} - \Delta G = \ell_p(e) - \ell_{p'}(e') + b[L(z_p) - L(z_p + e)]. \tag{J.4}$$

Finally, integrating $L'(v) = -(1 - F(v))$ gives $L(z_p) - L(z_p + e) = \tau(p)e - A_p(e)$. Substituting this and the quadratic bounds for $\ell_p(e)$ and $\ell_{p'}(e')$ proves Eq. (J.2) without requiring a differentiable density. □

For price-evaluation accuracy ε, SP permits inventory errors of order $\sqrt{\varepsilon}$, supporting the sharing in Lemma 6. For fixed $p < p'$ and $e' = 0$, CWZ2 generally requires $e = O(\varepsilon)$. Sampling and context-localization errors are separate. This comparison is specific to CWZ2: in the $a' > a$ branch, CWZ's Algorithm 5 has extra error $b(a' - a)[F(a) - F(z_p)]$, whose coefficient also shrinks with the truncation-point separation.

J.2 Controlled Experiments

Population comparison. Use $D(p) = c - p + \xi$, $\xi \sim \text{Unif}[-1, 1]$, $c = 3 + 1/12 + 0.005$, $b = h_c = 1$, and $\rho = 0.2$. The price range $[0.8, 2.05]$ and inventory range $[0, 4.5]$ keep demand nonnegative and the relevant quantiles interior. Compare $p = 1, p' = 2$, with $e' = 0$ and $|e| \le 0.04$. The optimized value is $G(p) = p(c - p) - (p + 1)/(p + 2)$, giving

$$\Delta G = 0.005, \qquad \Delta_{\text{SP}} - \Delta G = \tfrac{3}{4}e^2, \qquad \Delta_{\text{C2}} - \Delta G = \tfrac{1}{3}e + \tfrac{1}{2}e^2. \tag{J.5}$$

At $e = -0.04$, the SP and CWZ2 estimates are 0.006200 and -0.007533: the latter reverses the ranking despite an inventory profit loss of only 0.0012. Figure 5 shows the different sensitivities without sampling noise.

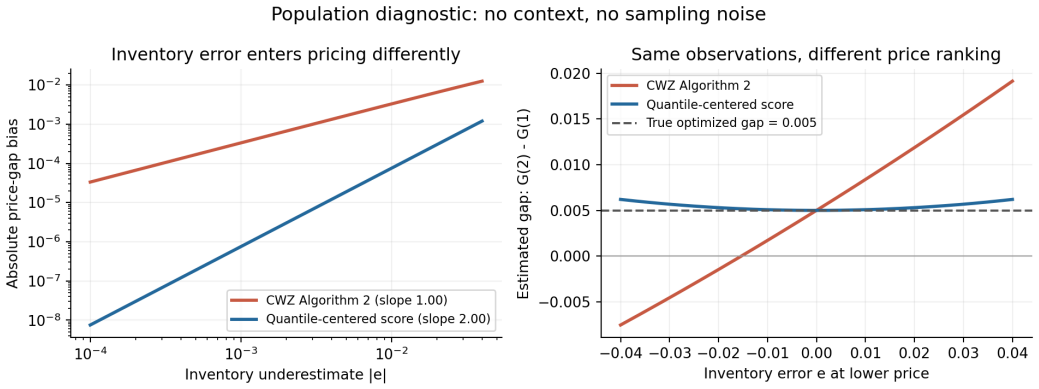


Figure 5: Population comparison with $e' = 0$. Left: comparison error scales linearly with $|e|$ for CWZ2 and quadratically for SP. Right: understocking can reverse CWZ2's price ranking. The true optimized-profit gap is 0.005.

Finite samples and a correction control. Keep the same demand form, costs, ranges, and candidate prices, and use $\xi = 2(U - \mathbb{E}U)$ with U distributed as $\text{Beta}(1, 1)$, $\text{Beta}(2, 2)$, or $\text{Beta}(2, 5)$. For each law, set $K_F(p) = (p + 2)L(z_p) + z_p$ and $c = 3 + K_F(2) - K_F(1) + 0.005$, preserving $\Delta G = 0.005$. Each price supplies $n \in \{1024, 4096, 16384, 65536, 262144\}$ independent observations, with 200 replications per law and sample size. All estimators use the same two independent price streams. We prescribe $e \in \{-0.04, 0, +0.04, -0.2n^{-1/4}\}$, always with $e' = 0$, and select the high price when the estimated gap is positive.

SP uses the empirical proxy mean and low sales quantile. CWZ2 uses its original supremum convention, the remaining-inventory order statistic with one-based index $\lfloor n\tau(p) \rfloor + 1$. A modified CWZ2 control replaces $\tau(p)$ by the actual empirical low-price stockout frequency in both inversion and correction. For $a' > a$, its population correction becomes $b[L(a) - L(a')]$, removing the additional error in Eq. (J.4). This control isolates the bias mechanism and differs from published CWZ2.

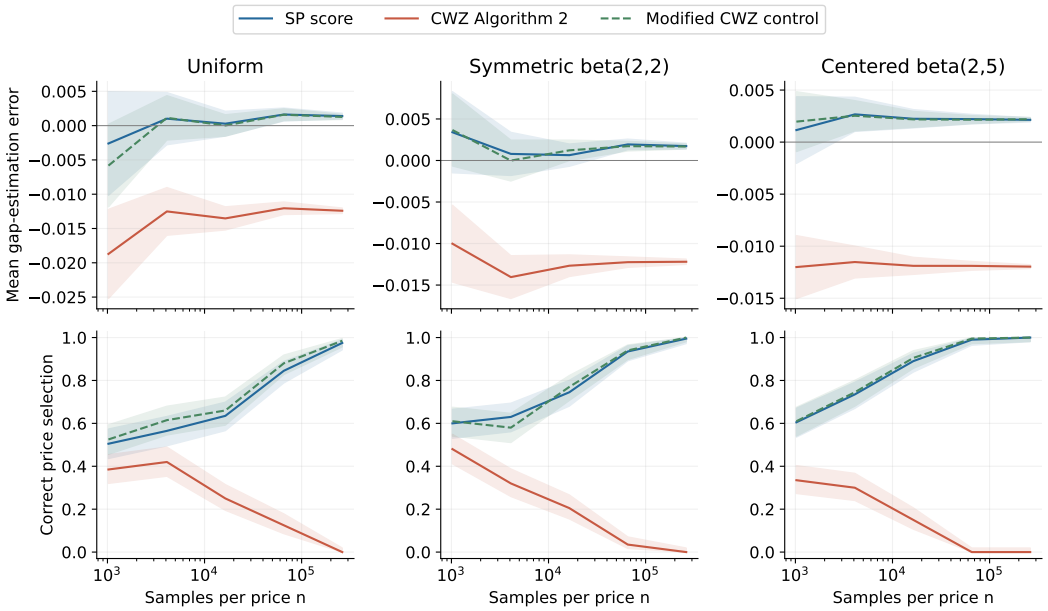


Figure 6: Fixed inventory error $e = -0.04, e' = 0$, using identical data across estimators. Top: mean gap error with 95% Monte Carlo intervals. Bottom: correct-price selection with 95% Wilson intervals over 200 replications. The modified control uses the actual stockout frequency.

Noise	Estimator	Population bias	Mean bias	Correct (%)
Uniform	SP score	+0.00120	+0.00139	97.5
Uniform	CWZ Algorithm 2	-0.01253	-0.01241	0.0
Uniform	Modified CWZ control	+0.00120	+0.00127	98.5
Beta(2,2)	SP score	+0.00172	+0.00173	99.5
Beta(2,2)	CWZ Algorithm 2	-0.01219	-0.01219	0.0
Beta(2,2)	Modified CWZ control	+0.00172	+0.00168	100.0
Beta(2,5)	SP score	+0.00235	+0.00214	100.0
Beta(2,5)	CWZ Algorithm 2	-0.01177	-0.01196	0.0
Beta(2,5)	Modified CWZ control	+0.00235	+0.00220	100.0

Table 1: Fixed understocking, $e = -0.04, e' = 0$, at $n = 262144$ observations per price. Population biases are exact. Empirical results use 200 replications.

Figure 6 and Table 1 show the finite-sample effect of the population bias under fixed understocking. As the sample size increases, CWZ2 concentrates around the biased price difference, and the ranking error persists as sampling noise vanishes. The modified control traces this bias to using the critical stockout probability in the correction at a misspecified inventory. Our visible low-quantile correction remains valid at that inventory, leaving only the quadratic inventory profit loss in the price comparison. This is the mechanism that allows coarse inventory estimates to support finer price learning.

K AI Use Statement

Generative AI tools assisted with editing and manuscript organization. The authors independently wrote the proofs and take responsibility for the manuscript.