Offline-to-Online VAPE: Conservative Contextual Dynamic Pricing from Biased Logs

Liz Lemma Future Detective

January 16, 2026

Abstract

We study contextual dynamic pricing with one-bit purchase feedback, building on the VAPE (Valuation Approximation–Price Elimination) framework that separates learning the contextual valuation function from learning a shared demand curve over price increments. While VAPE attains minimax-optimal $\tilde{O}(T^{2/3})$ regret under minimal assumptions, it relies on deliberately randomized prices to obtain unbiased valuation signals—often infeasible in modern (2026) marketplaces with experimentation governance and customer-trust constraints. We propose Offline-to-Online VAPE: an algorithm that uses historical logged pricing data to warm-start both valuation and demand estimation, thereby reducing the frequency of disruptive random-price rounds while preserving the $\tilde{O}(T^{2/3})$ asymptotic regret rate in the linear valuation model with bounded noise and Lipschitz demand. Our main analysis quantifies how overlap in the logging policy (a small randomized component) translates into fewer valuation-approximation rounds via an improved elliptical-potential argument from a stronger initial design matrix. We further provide a conservative wrapper that guarantees high-probability safe improvement: cumulative online revenue stays above a certified fraction of a baseline policy value. The resulting framework connects modern offline evaluation (IPS/DR) with online learning-to-price under minimal structural assumptions.

Table of Contents

3. 3. Logged data and overlap: define the logging policy class; identifiability/overlap conditions; what can and cannot be learned from biased logs; discussion of realistic logging (small randomized exploration).

4. 4. Offline initialization: ridge-style estimation of θ from randomized log slices; construction of initial increment-demand estimates $\hat{D}_0(\delta_k)$; DR/IPS evaluation of baseline value R_0 and a lower confidence bound $\underline{R}_0$.

5. 5. Offline-to-Online VAPE algorithm: initialization + online VAPE with prior counts; optional conservative wrapper (budget-based mixing with baseline). Pseudocode and computational complexity.

6. 6. Main theory results: (i) high-probability offline estimation bounds; (ii) reduced number of valuation-approximation rounds via log-det improvement; (iii) regret bound combining offline error + online learning; (iv) baseline-safety guarantee.

7. 7. Comparative statics and design guidance: how overlap ρ, log size n, and grid precision ε change exploration burden, regret constants, and safety slack; practical tuning recommendations.

8. 8. Simulations / empirical template (optional for workshop): synthetic drift-free and mild-drift settings; illustrate exploration reduction; ablations (with/without conservative wrapper; varying ρ).

9. 9. Limitations and extensions: unknown propensities, weak overlap, nonstationary logs vs online, context-dependent noise; where numerical methods would be needed (e.g., estimating propensities, fitting flexible nuisance models for DR).

1 Introduction

A recurring friction in deployed dynamic pricing systems is that the learning signal is often *binary*—we observe whether a buyer purchases at a posted price, but not the buyer's valuation. This informational constraint turns online experimentation into a delicate exercise: to learn where demand bends, we must occasionally post prices that are "wrong" for the current context, and those deviations can be *disruptive*. They are disruptive in the narrow sense of foregone short-run revenue, and in the broader sense of product and policy risk: extreme prices can degrade user trust, trigger complaints, or violate internal governance rules that require stability relative to a vetted baseline. In our setting, the central bottleneck is therefore not merely statistical efficiency, but the *economics of exploration*: how to acquire incremental information while limiting the economic and operational cost of experimentation.

The VAPE approach (valuation approximation and price elimination) offers a useful lens on this bottleneck. At a high level, VAPE recognizes that with only purchase/no-purchase feedback, learning demand at a given context is hard unless we can translate that binary feedback into an approximate valuation signal. Doing so typically requires posting randomized prices on a sufficiently rich support, which is exactly what creates disruption. Once a valuation proxy is available, the algorithm can reduce uncertainty about the parameter governing the valuation function and, in parallel, learn a one-dimensional object capturing how purchase probabilities shift with price increments. The decomposition is conceptually appealing: we pay a "tax" during a limited set of valuation-approximation rounds, and then exploit the learned structure to narrow down the optimal price among a discretized set of candidates. Yet in the cold-start regime, the early valuation-approximation phase can be long enough to dominate realized performance, especially when the context dimension is moderate and the platform is unwilling to tolerate large price swings.

Our main motivation is that many platforms are not truly in a cold-start regime. Prior to launching an adaptive policy, firms frequently possess *offline logs* generated by a previously deployed baseline pricing rule, sometimes augmented by deliberate randomization for experimentation or compliance. These logs are usually collected for business analytics, A/B testing, or auditing, but they carry precisely the kind of overlap that valuation-approximation methods require: occasional uniform or wide-support random prices that can be treated as exogenous conditional on context. From an economic perspective, the existence of such logged randomization means that some of the "costly" exploration has already been paid in the past. It is therefore natural to ask how we can convert this historical experimentation into a *warm start* that reduces disruption during future online learning, without changing the online objective or weakening regret guarantees.

This paper develops an offline-to-online instantiation of VAPE that does exactly that. The idea is simple: we use the portion of the logs generated by the randomized component of the logging policy to initialize (i) a linear estimator of the valuation parameter and (ii) initial estimates of purchase probabilities at a grid of price increments. These two initializations are complementary. The valuation warm start reduces the need for early online rounds devoted to constructing valuation proxies, while the increment-demand warm start shrinks confidence intervals for demand at relevant price differences, accelerating elimination of dominated prices. The result is an algorithm that retains the same asymptotic regret exponent as VAPE, but with a smaller leading constant governed by an effective offline information quantity (captured by the initial design matrix). Put differently, the logs do not change the fundamental difficulty of learning from binary feedback, but they can substantially reduce the *time spent in the disruptive regime*.

A second motivation is governance. In practice, an adaptive pricing policy is rarely deployed without guardrails relative to an approved baseline. These guardrails may be motivated by revenue risk (e.g., "do not underperform the incumbent policy by more than a fixed margin"), by customer protection (avoid unexpectedly high prices), or by organizational accountability (ensure that experimentation can be justified ex post). We therefore frame an optional conservative wrapper around the learning policy: online decisions are permitted to deviate from the baseline only when the algorithm's current evidence suggests that doing so is safe, and otherwise the baseline is played. The conceptual analogy is to a budgeting rule: past gains relative to a conservative lower bound can finance future experimentation, while preventing "bankruptcy" relative to the benchmark. This wrapper separates two concerns that are often conflated in discussions of safe learning: estimating the demand/valuation model efficiently, and controlling realized performance relative to a reference policy.

Several limitations are worth stating up front. Our ability to reuse logs hinges on *known propensities* and *overlap*: if the historical data contain too little randomization, or if logging probabilities are misspecified, then offline estimates can be biased and the warm start can be misleading. Likewise, the high-probability guarantees we provide rely on standard regularity conditions (boundedness and mild smoothness of the noise distribution) that discipline how sharply demand can change with price. These assumptions are not innocuous, but they match the operational reality that platforms typically impose explicit price bounds and monitor volatility. Finally, our safety wrapper protects revenue relative to a baseline lower confidence bound; it does not, by itself, address broader normative constraints (such as fairness across groups) without additional structure.

With these motivations in place, we proceed as follows. The next section formalizes the online pricing problem with binary feedback, specifies the structural assumptions that make learning feasible, and reviews the VAPE

decomposition that underpins our algorithmic design.

2 Baseline online model and the VAPE decomposition

We begin with the online problem absent any historical data. Time is indexed by $t = 1, \ldots, T$. At each round the seller observes a context vector $x_t \in \mathbb{R}^d$ (features describing the buyer, product, or market state) satisfying $\|x_t\|_2 \leq B_x$, and then posts a price $p_t \in [0, B_y]$. The buyer has a latent valuation

$$y_t \;=\; x_t^\top \theta + \xi_t,$$

where $\theta \in \mathbb{R}^d$ is unknown and bounded as $\|\theta\|_2 \leq B_\theta$. The noise ξ_t is i.i.d., centered, and bounded $|\xi_t| \leq B_\xi$. We assume its CDF F is L_ξ-Lipschitz, which implies a mild smoothness of purchase probabilities in price. The seller observes only a binary outcome

$$o_t \;=\; \mathbf{1}\{p_t \leq y_t\},$$

and earns revenue $r_t = p_t o_t$. This is the canonical friction in many pricing deployments: we do not see y_t, we see only whether the posted price clears the buyer's willingness-to-pay.

The model implies a convenient factorization of expected revenue. Define the *demand-increment* (or tail) function

$$D(\delta) \;=\; \mathbb{P}(\xi \geq \delta) \;=\; 1 - F(\delta).$$

Then conditional on (x, p) we have

$$\pi(x, p) \;=\; \mathbb{E}\Big[p\,\mathbf{1}\{p \leq x^\top \theta + \xi\} \mid x, p\Big] \;=\; p\,D\Big(p - x^\top \theta\Big)\,.$$

Economically, $p - x^\top \theta$ is the *price increment* above the context-dependent mean valuation $g(x) = x^\top \theta$, and $D(\cdot)$ maps that increment into a purchase probability. The Lipschitz assumption on F equivalently gives $|D(\delta) - D(\delta')| \leq L_\xi |\delta - \delta'|$, ruling out infinitely sharp discontinuities in demand as we vary price.

Given a realized context sequence $\{x_t\}_{t=1}^T$, the benchmark is the best price in hindsight at each round, i.e.,

$$p^*(x) \in \arg\max_{p \in [0, B_y]} \; p\,D(p - x^\top \theta).$$

We evaluate an online policy $\{p_t\}$ via (pseudo-)regret

$$R_T \;=\; \sum_{t=1}^{T} \max_{p \in [0, B_y]} \pi(x_t, p) - \sum_{t=1}^{T} \pi(x_t, p_t) \;=\; \sum_{t=1}^{T} \max_{p \in [0, B_y]} p\,D(p - x_t^\top \theta) - \sum_{t=1}^{T} p_t\,D(p_t - x_t^\top \theta).$$

This objective makes the central tradeoff transparent. To earn revenue we want p_t near $p^*(x_t)$, but to learn $p^*(\cdot)$ we must infer both the *level* $x^\top\theta$ and the *shape* of $D(\cdot)$ using only binary outcomes.

The VAPE principle (*valuation approximation and price elimination*) is to separate these two learning tasks. The key observation is that the expected reward depends on the context only through $x^\top\theta$, while all residual price-response is captured by the one-dimensional function $D(\delta)$. VAPE leverages this structure by alternating between: (i) acquiring an approximate valuation signal to estimate θ, and (ii) learning $D(\cdot)$ on a discretized grid of increments to eliminate suboptimal prices.

The first component is the *valuation approximation* step. Binary feedback becomes informative about θ only if, conditional on x_t, we sometimes post prices on a sufficiently rich support. Under a symmetric uniform randomization over an interval containing the feasible prices (conceptually, $p_t \sim \mathrm{Unif}([-B_y, B_y])$), one can construct a pseudo-outcome

$$z_t \;=\; 2B_y\Big(o_t - \tfrac{1}{2}\Big)$$

that satisfies an unbiasedness identity $\mathbb{E}[z_t \mid x_t] = x_t^\top\theta$ (the intuition is that, under uniform pricing, the purchase indicator integrates the valuation threshold in a way that recovers the mean). This converts the pricing problem into a linear regression with bounded noise, so that standard self-normalized concentration controls $\|\hat\theta - \theta\|_V$ for a design matrix $V = I + \sum x_t x_t^\top$. Importantly, these valuation-approximation rounds are precisely the ones that are operationally disruptive: they deliberately inject price variation to obtain identification.

The second component is *increment-demand learning* via discretization. Fix a grid resolution $\varepsilon > 0$ and grid points $\delta_k = k\varepsilon$ for $k \in \mathcal{K}$ with $|\delta_k| \le B_y$. For a given context x, if we had a good estimate $\hat\theta$ then we could translate a candidate price p into an estimated increment $\hat\delta = p - x^\top\hat\theta$ and hence into a nearby grid point δ_k. VAPE maintains empirical estimates of $D(\delta_k)$ from observed purchase outcomes at prices whose implied increments fall in bin k, along with confidence intervals that shrink at the usual $\sqrt{1/N_k}$ rate. The Lipschitz property of D ensures that binning incurs only $O(\varepsilon)$ approximation error.

Operationally, VAPE uses these ingredients to eliminate prices that are provably dominated given current confidence sets for (θ, D). One can think of the algorithm as maintaining a set of plausible demand curves and valuation parameters; if a candidate increment δ_k cannot be optimal for any plausible model, it is dropped. Over time, the remaining candidate set concentrates around the revenue-maximizing increment, and pricing becomes less exploratory.

Two limitations of this baseline picture are worth flagging. First, without any prior information, early valuation-approximation can consume a

nontrivial fraction of the horizon because uncertainty about θ is high in d dimensions. Second, the uniform randomization that makes valuation approximation clean is exactly the type of behavior platforms often wish to minimize. These frictions motivate our next step: exploiting logged overlap to shift part of the disruptive experimentation from the online phase into an offline warm start.

3 Logged data and overlap: what historical prices do (and do not) identify

We now introduce the offline log and make explicit the role of *overlap*. Before online interaction begins, the seller observes a dataset

$$\mathcal{L} = \{(x_i^L, p_i^L, o_i^L)\}_{i=1}^n,$$

generated by the same valuation model as online: conditional on x_i^L, the buyer valuation is $y_i^L = (x_i^L)^\top \theta + \xi_i$ with the same unknown θ and the same i.i.d. noise distribution for ξ_i. The key difference relative to the online phase is that prices in the log are not chosen by our learning algorithm; they are assigned by a historical *logging policy* (typically a production heuristic, possibly with small randomized perturbations).

To reflect the operational reality that many platforms run limited exploration (e.g., occasional A/B perturbations for measurement), we assume the logging propensities are known and take the mixture form

$$\mu_0(p \mid x) = \rho \,\mathrm{Unif}([-B_y, B_y]) + (1-\rho)\,\delta_{\pi_0(x)},$$

where $\pi_0(x)$ is a baseline pricing rule and $\rho \in (0, 1]$ is the logged exploration overlap. This class captures two stylized facts: (i) most of the time the system plays a deterministic baseline price $\pi_0(x)$ chosen for revenue or business constraints, and (ii) with small probability ρ the system deviates to a price drawn from a broad support. The symmetry of $\mathrm{Unif}([-B_y, B_y])$ is analytically convenient for valuation approximation; in practice, one can interpret the negative part as a normalization device (or as allowing rebates/credits), and our bounds depend primarily on the existence of a known distribution with sufficiently wide support.

The central identification issue is that, with binary outcomes, *variation in prices conditional on context* is what turns purchase data into information about $x^\top\theta$ and about the tail function $D(\cdot)$. If the logger were purely deterministic ($\rho = 0$), then conditional on a fixed x we observe only $o = \mathbf{1}\{\pi_0(x) \le x^\top\theta + \xi\}$, i.e., a single threshold event at one price. Such data can be informative for predicting purchase under the *same* policy π_0—it directly identifies the purchase probability at that posted price—but it does not, by itself, identify counterfactual purchase probabilities at other prices,

nor does it cleanly identify the linear index $x^\top\theta$. Put differently, deterministic logging leads to a form of set identification: many combinations of (θ, F) can rationalize the same mapping $x \mapsto \mathbb{P}(o = 1 \mid x, p = \pi_0(x))$, because we never observe how demand changes when price is perturbed around $\pi_0(x)$.

Even when we maintain the linear structure $g(x) = x^\top\theta$, biased logs can still be problematic because the baseline price $\pi_0(x)$ is typically chosen as a function of x (and possibly other unobserved state). This adaptivity induces a strong correlation between the chosen price and the latent valuation component $x^\top\theta$, so naive regressions of o on x or p are generally invalid for recovering θ or D. The correct object we need for learning is not $\mathbb{E}[o \mid x]$ under the baseline, but rather $\mathbb{P}(o = 1 \mid x, p)$ as a function of p at fixed x, which requires overlap in the conditional price distribution.

This is precisely what $\rho > 0$ supplies. On rounds where the logger uses the uniform component, the price is independent of ξ conditional on x, and moreover has a known, rich support. That combination delivers two benefits. First, it yields an unbiased valuation signal: on uniformly randomized log rounds one can construct the pseudo-outcome $z = 2B_y(o - \frac{1}{2})$ satisfying $\mathbb{E}[z \mid x] = x^\top\theta$. This converts a subset of the logged data into a standard linear estimation problem for θ, with effective sample size on the order of ρn. Second, broad price support implies broad support over increments $\delta = p - x^\top\theta$, which is what we need to learn the one-dimensional demand-increment curve $D(\delta) = \mathbb{P}(\xi \geq \delta)$ beyond the narrow region induced by the baseline policy.

The notion of *overlap* is therefore not an abstract technicality; it is the formal expression of the business practice of running small but systematic exploration to make future improvements possible. The mixture model makes the tradeoff stark. As ρ increases, we get more randomized data and hence tighter offline confidence sets; but the logger departs more often from the baseline policy. As ρ decreases, the log becomes safer and closer to business-as-usual, but offline learning becomes weaker, and at the limit $\rho \downarrow 0$ we cannot hope to initialize the online learner in a statistically reliable way.

Finally, we emphasize what the log can and cannot buy us. Logged data can reduce online disruption only to the extent that it contains genuine experimental variation with known propensities. It cannot, by itself, eliminate the need for online learning, because the contexts $\{x_t\}$ arriving online may differ from those in the log, and because binary feedback fundamentally limits how quickly we can localize the optimal increment without continued data collection. Moreover, if the propensities $\mu_0(p \mid x)$ are misspecified or unobserved, off-policy reasoning becomes fragile: without correct propensities, even evaluating the baseline reliably can fail, undermining any safety guarantee built from a lower confidence bound. With these caveats in place, the mixture logger provides a clean and realistic route to offline-to-online transfer: it supplies a tractable set of "as-if randomized" rounds that we can use to warm start estimation while preserving the operational interpretation

of small randomized exploration.

4 Offline initialization: learning θ, seeding demand increments, and certifying the baseline

We now describe how we use the log $\mathcal{L}$ to warm start the online VAPE routine. Conceptually, we want three offline objects: (i) an initial confidence set for the linear index $x^\top\theta$ (implemented via a ridge-style estimator and design matrix), (ii) initial counts and empirical estimates for the one-dimensional demand-increment curve $D(\delta) = \mathbb{P}(\xi \geq \delta)$ on a discretized grid, and (iii) an offline lower confidence bound on the baseline value, which will later power an optional conservative (baseline-safe) wrapper.

Step 1: Ridge-style estimation of θ from randomized log slices. Because binary feedback makes direct regression of o on (x, p) ill-posed under adaptive logging, we deliberately restrict attention to the "as-if randomized" subset of the log. Let $\mathcal{I}_u \subset \{1, \ldots, n\}$ be the indices for which the logger drew p_i^L from the uniform component of $\mu_0(\cdot \mid x_i^L)$. Operationally, this is a flag that many systems can record by construction (and it is statistically the cleanest source of identification).

On these rounds we form the VAPE pseudo-outcome

$$z_i \;=\; 2B_y\Big(o_i^L - \tfrac{1}{2}\Big), \qquad i \in \mathcal{I}_u,$$

which satisfies the key identity $\mathbb{E}[z_i \mid x_i^L] = (x_i^L)^\top\theta$ under uniform random pricing. We then run a ridge regression on (x_i^L, z_i):

$$V_0 \;=\; I + \sum_{i\in\mathcal{I}_u} x_i^L (x_i^L)^\top, \qquad \hat{\theta}_0 \;=\; V_0^{-1} \sum_{i\in\mathcal{I}_u} x_i^L z_i.$$

Two features are worth emphasizing. First, the effective sample size is $|\mathcal{I}_u| \approx \rho n$, so overlap directly controls the tightness of the warm start. Second, V_0 is not merely a computational artifact: it is the initial "geometry" that will govern online uncertainty via norms of the form $\|x\|_{V_0^{-1}}$. In particular, directions in feature space that were well-covered by logged randomization will be treated as less uncertain online.

Step 2: Seeding increment-demand estimates on a price grid. VAPE learns demand through increments $\delta = p - x^\top\theta$, so we discretize the increment space with a grid $\{\delta_k\}_{k\in\mathcal{K}}$, where $\delta_k = k\varepsilon$ and ε will match the online resolution. Using $\hat{\theta}_0$, we map each randomized log observation to an estimated increment

$$\hat{\delta}_i \;=\; p_i^L - (x_i^L)^\top\hat{\theta}_0, \qquad i \in \mathcal{I}_u,$$

and assign $\hat{\delta}_i$ to its nearest grid point δ_k (ties broken arbitrarily). For each bin k we define the offline count and purchase-rate estimate

$$N_{k,0} \;=\; \sum_{i\in\mathcal{I}_u} \mathbf{1}\{\hat{\delta}_i \mapsto \delta_k\}, \qquad \hat{D}_{k,0} \;=\; \frac{1}{N_{k,0}} \sum_{i\in\mathcal{I}_u} \mathbf{1}\{\hat{\delta}_i \mapsto \delta_k\}\, o_i^L,$$

whenever $N_{k,0} \geq 1$. This produces a warm-started estimate of $D(\delta_k)$ for all increments that are sufficiently represented in the randomized slice.

The economic logic is simple: conditional on x, purchase is exactly the event $\{\xi \geq p - x^\top\theta\}$, so once we can place observations on the increment axis, learning demand becomes a one-dimensional nonparametric estimation problem. Statistically, two biases enter and are controlled by our assumptions: discretization creates an $O(\varepsilon)$ approximation error, and using $\hat{\theta}_0$ instead of θ shifts increments by $(x^\top(\hat{\theta}_0 - \theta))$, which translates into demand error via Lipschitzness of D. These are precisely the terms we later carry into confidence radii for elimination in the online phase.

Step 3: Offline evaluation of the baseline and a lower confidence bound. If we want baseline safety online, we need a conservative benchmark for the baseline per-round value

$$R_0 \;=\; \mathbb{E}_x\big[\pi(x, \pi_0(x))\big] \;=\; \mathbb{E}_x\big[\pi_0(x)\, D(\pi_0(x) - x^\top\theta)\big].$$

Because μ_0 is known, we can estimate R_0 using standard off-policy tools. A particularly transparent construction is a doubly robust (DR) estimator built from (i) an importance weight that selects rounds on which the logger actually played the baseline price and (ii) a plug-in reward model $\hat{\pi}(x, p)$ derived from $(\hat{\theta}_0, \hat{D}_{k,0})$. Concretely, define

$$\hat{\pi}(x, p) \;=\; p\, \hat{D}_0(p - x^\top\hat{\theta}_0),$$

where $\hat{D}_0(\cdot)$ denotes the grid-based interpolation induced by $\{\hat{D}_{k,0}\}$. For each log row, set the baseline indicator $a_i = \mathbf{1}\{p_i^L = \pi_0(x_i^L)\}$ (or, in implementation, equality up to a known pricing tick), and define the propensity of that action under the mixture logger as $\mu_0(\pi_0(x_i^L) \mid x_i^L) = 1 - \rho$ (the uniform component contributes no point mass). The DR estimate is then

$$\hat{R}_{\mathrm{DR}} \;=\; \frac{1}{n}\sum_{i=1}^{n} \Big(\hat{\pi}(x_i^L, \pi_0(x_i^L)) + \frac{a_i}{1-\rho}\big(r_i^L - \hat{\pi}(x_i^L, p_i^L)\big)\Big), \qquad r_i^L = p_i^L o_i^L \in [0, B_y].$$

We then form a scalar lower confidence bound

$$\underline{R}_0 \;=\; \hat{R}_{\mathrm{DR}} - \mathrm{rad}(n, \delta),$$

where $\mathrm{rad}(n, \delta)$ is chosen via a bounded-difference concentration inequality (e.g., Hoeffding or an empirical Bernstein bound) to ensure $\mathbb{P}(R_0 \geq \underline{R}_0) \geq$

$1-\delta/2$. The practical point is that, for safety, we only need a *one-sided* guarantee on a *single number*. This is far less demanding than fully learning $D(\cdot)$ everywhere, but it does hinge on correct propensities; if μ_0 is misspecified, $\underline{R}_0$ may be over-optimistic unless we add robustness or sensitivity adjustments.

Taken together, $(V_0, \hat{\theta}_0, \{N_{k,0}, \hat{D}_{k,0}\}_{k\in\mathcal{K}}, \underline{R}_0)$ summarize what the log contributes to the online phase: less initial uncertainty in high-coverage directions, nontrivial prior counts that reduce early demand-learning variance, and (optionally) a certified baseline floor that can be enforced by a conservative wrapper.

5 Offline-to-Online VAPE: warm start, prior counts, and an optional conservative wrapper

We now describe the actual offline-to-online procedure that consumes the offline summary

$$\Big(V_0, \hat{\theta}_0, \{N_{k,0}, \hat{D}_{k,0}\}_{k\in\mathcal{K}}, \underline{R}_0\Big)$$

and produces online prices $\{p_t\}_{t=1}^{T}$. The goal is to preserve VAPE's basic logic—separating the (high-dimensional) problem of tracking $x^\top\theta$ from the (one-dimensional) problem of learning $D(\cdot)$—while using logs to reduce early uncertainty. The only substantive change relative to a cold start is that we treat the offline quantities as "pseudo-observations" at time $t = 0$: the online design begins at V_0 rather than I, the online ridge estimate begins at $\hat{\theta}_0$, and each increment bin begins with count $N_{k,0}$ and mean $\hat{D}_{k,0}$.

Online state variables. Online we maintain (i) a ridge state for the linear index and (ii) binned statistics for increment demand. Concretely, we keep a design matrix V_t and score vector b_t such that $\hat{\theta}_t = V_t^{-1}b_t$, initialized at V_0 and $b_0 = \sum_{i\in\mathcal{I}_u} x_i^I z_i$ so that $\hat{\theta}_0 = V_0^{-1}b_0$. In parallel, for each increment index $k \in \mathcal{K}$ we keep an online count $N_{k,t}$ and empirical mean $\hat{D}_{k,t}$, initialized at $N_{k,0}$ and $\hat{D}_{k,0}$ (and interpreted as a prior mean computed from data, not a Bayesian prior).

To map increments into prices we use the projection operator $\Pi_{[0,B_y]}(u) = \min\{B_y, \max\{0, u\}\}$, and define the candidate price associated with increment δ_k as

$$p_{t,k} \;=\; \Pi_{[0,B_y]}\big(x_t^\top\hat{\theta}_{t-1} + \delta_k\big).$$

Given $N_{k,0} + N_{k,t-1}$ observations in bin k, we form a confidence interval for $D(\delta_k)$ using a Hoeffding-style radius (augmented by the same ε-discretization and index-shift terms that appear in the offline analysis). We denote generic upper and lower bounds by

$$\mathrm{UCB}_{k,t} \in [0,1], \qquad \mathrm{LCB}_{k,t} \in [0,1],$$

with $\mathrm{LCB}_{k,t} \le D(\delta_k) \le \mathrm{UCB}_{k,t}$ on the high-probability event. The warm start enters only through the effective sample size $N_{k,0} + N_{k,t-1}$.

When do we "query" valuation? VAPE does not update the linear index on every round; instead, it triggers a valuation-approximation step only when the current context is insufficiently covered by the design. A convenient rule is to define a threshold $\mu > 0$ and declare round t a valuation round if

$$\|x_t\|_{V_{t-1}^{-1}} > \mu.$$

On valuation rounds we post a uniformly random price (or any symmetric randomization with the VAPE identity), observe o_t, and update the ridge state using the pseudo-outcome $z_t = 2B_y(o_t - \frac{1}{2})$. On non-valuation rounds we treat $\hat{\theta}_{t-1}$ as accurate enough for purposes of choosing among increment candidates, and we focus learning effort on $D(\cdot)$.

Warm-started online VAPE (with optional baseline safety). The full procedure is summarized below.

Input: $V_0, \hat{\theta}_0, \{N_{k,0}, \hat{D}_{k,0}\}_{k\in\mathcal{K}}$, grid $\{\delta_k\}$, confidence α, threshold μ.
Optional input: baseline π_0, LCB $\underline{R}_0$, slack β, initial budget $B_1 \ge 0$.

Initialize $V \leftarrow V_0$, $b \leftarrow V_0\hat{\theta}_0$; set $N_k \leftarrow N_{k,0}$ and $\hat{D}_k \leftarrow \hat{D}_{k,0}$ for all k.
For $t = 1, \dots, T$:
  Observe x_t.
  (Safety wrapper, optional) if $B_t < 0$ **then** set $p_t \leftarrow \pi_0(x_t)$ and go to "Update budget".
  (Valuation approximation) if $\|x_t\|_{V^{-1}} > \mu$ **then**
    Draw $p_t \sim \mathrm{Unif}([-B_y, B_y])$, post $\Pi_{[0,B_y]}(p_t)$, observe o_t.
    Set $z_t = 2B_y(o_t - \frac{1}{2})$; update $V \leftarrow V + x_t x_t^\top$, $b \leftarrow b + x_t z_t$.
    Set $\hat{\theta} \leftarrow V^{-1}b$.
  else (Demand learning / exploitation):
    For each $k \in \mathcal{K}$, compute $p_{t,k} = \Pi_{[0,B_y]}(x_t^\top\hat{\theta} + \delta_k)$ and an optimistic value $\hat{\pi}_{t,k} = p_{t,k} \cdot \mathrm{UCB}$
    Choose $k_t \in \arg\max_k \hat{\pi}_{t,k}$, set $p_t \leftarrow p_{t,k_t}$, observe o_t.
    Update bin statistics for k_t: $N_{k_t} \leftarrow N_{k_t} + 1$, $\hat{D}_{k_t} \leftarrow \frac{(N_{k_t}-1)\hat{D}_{k_t} + o_t}{N_{k_t}}$.
  Update budget (optional): set $r_t = p_t o_t$ and $B_{t+1} \leftarrow B_t + r_t - (1-\beta)\underline{R}_0$.
End for.

Two remarks clarify the economic role of the safety wrapper. First, the budget recursion enforces an intertemporal participation constraint relative to the offline-certified baseline floor: when the algorithm has "banked" enough surplus, it can experiment; when it falls behind, it temporarily reverts to π_0. Second, the wrapper is modular: it does not change how we compute candidate prices or confidence bounds, only when we are allowed to deploy them.

Computational complexity. The dominant per-round cost is scanning the increment grid. Since $|\mathcal{K}| \asymp B_y/\varepsilon$, computing $\{p_{t,k}, \hat{\pi}_{t,k}\}_{k\in\mathcal{K}}$ costs $O(|\mathcal{K}|)$ arithmetic operations per non-valuation round, i.e.,

$$O\left(\frac{T}{\varepsilon}\right) \quad \text{overall,}$$

up to constant factors from projection and confidence-radius evaluation. Ridge updates are needed only on valuation rounds. Using Sherman–Morrison, each update of V^{-1} costs $O(d^2)$, so the total linear-algebra cost is

$$O\big(|\mathcal{G}_{on}|\, d^2\big),$$

where $\mathcal{G}_{on}$ is the (random) set of valuation-approximation rounds. Storage is $O(d^2 + |\mathcal{K}|)$ for V^{-1} (or a Cholesky factor) plus the per-bin counts and means. In particular, the offline warm start reduces computation in the same place it reduces regret: by shrinking $|\mathcal{G}_{on}|$ through a larger initial design V_0.

6 Comparative statics and design guidance

Our theory highlights a simple economic logic: logged randomization buys us *credible* early information about the valuation index and the demand curve, and that information substitutes for disruptive online experimentation. The comparative statics therefore run through two summary objects—the initial design volume $\det(V_0)$ for the high-dimensional index, and the initial bin counts $\{N_{k,0}\}_{k\in\mathcal{K}}$ for the one-dimensional increment demand.

Overlap ρ: why even small randomization matters. When $\mu_0(p \mid x) = \rho\,\mathrm{Unif}([-B_y, B_y]) + (1-\rho)\delta_{\pi_0(x)}$ has $\rho > 0$, the uniform component produces unbiased VAPE signals and hence a genuine linear-regression "sample size" of order $|\mathcal{I}_u| \approx \rho n$. This affects online learning through (i) a smaller initial estimation radius for θ (Proposition 1) and (ii) a larger initial determinant $\det(V_0)$, which shrinks the bound on valuation-approximation rounds (Proposition 3). A convenient back-of-the-envelope relation is

$$\log\det(V_0) \;\approx\; \sum_{j=1}^{d} \log\big(1+\lambda_j\big) \;\lesssim\; d\log\Big(1+\tfrac{B_x^2\,\rho n}{d}\Big),$$

where $\{\lambda_j\}$ are eigenvalues of $\sum_{i\in\mathcal{I}_u} x_i^L (x_i^L)^\top$. Because the dependence is logarithmic, we get diminishing returns: increasing ρ from 0 to a small positive value can be qualitatively important (it restores identifiability and warm-start feasibility), while pushing ρ from, say, moderate to large yields more modest improvements.

In practice, the policy implication is that if we control logging, it is often worth paying a small short-run revenue cost to ensure nontrivial randomization, since it reduces the future need for online random prices that are typically more salient (and potentially more costly) than offline experimentation.

Log size n: more data helps, but with diminishing returns. Holding ρ fixed, increasing n raises both $\det(V_0)$ and the demand-bin counts $N_{k,0}$. The first reduces the frequency of valuation rounds; the second tightens confidence intervals for $D(\delta_k)$ immediately, which reduces the "demand-elimination" component of regret and accelerates reliable exploitation. The diminishing-returns logic is again important: because the valuation-side benefit enters via $\log\det(V_0)$, doubling n does not halve experimentation, but it can still noticeably reduce the early, high-variance phase when V_t is small. This suggests a clear operational guidance: if one anticipates a short online horizon T, investing in a larger offline log can be particularly valuable, because it front-loads information that would otherwise need to be acquired online.

Grid precision ε: a bias–variance tradeoff with an offline "variance subsidy." The VAPE decomposition makes the role of ε transparent: finer grids reduce discretization loss but increase the statistical burden of learning demand at many increments. In the initialized regret bound, the terms

$$T\varepsilon \qquad \text{and} \qquad \varepsilon^{-2}\log(T)$$

represent, respectively, discretization and learning complexity (up to constants and lower-order logarithms), while the warm start primarily reduces the constant factors inside the demand-confidence radii through $N_{k,0}$. Economically, the offline counts act like a variance subsidy: when $N_{k,0}$ is large for the increments that matter, we can afford a somewhat finer ε (and hence less price-discretization bias) without paying as much online exploration.

A practical tuning rule is therefore:

- Use the theoretical default $\varepsilon = (d^2\log^2 T/T)^{1/3}$ as a safe baseline.
- If logs are large and well-spread so that many bins have substantial $N_{k,0}$, consider decreasing ε until the smallest "relevant" bins still have enough effective samples (offline + expected online) to keep Hoeffding radii small.
- If logs are sparse in the tails of δ, avoid overly fine uniform grids; instead, truncate $\mathcal{K}$ to increments that are empirically populated, or use coarser bins in regions with few observations (at the cost of local discretization bias).

This is also where we should acknowledge a limitation: the binning analysis uses Lipschitzness of F to translate increment errors into demand errors; if $D(\cdot)$ has sharp changes, the practical value of very fine grids may be constrained by model misspecification rather than sample size.

Safety slack β: calibrating conservatism to evaluation quality. The conservative wrapper guarantees (with high probability) that cumulative revenue does not fall below $(1-\beta)T\,\underline{R}_0$. Comparative statics are immediate: a larger β weakens the required floor and therefore reduces forced-baseline plays, typically improving regret. But the economically relevant object is not β alone; it is β *relative to the tightness of* $\underline{R}_0$. If offline evaluation is sharp (good overlap, correct propensities, stable context distribution), then $\underline{R}_0$ is close to R_0 and we can choose a smaller β while still permitting meaningful online learning. Conversely, if propensity knowledge is uncertain or contexts drift between log and deployment, then $\underline{R}_0$ may be conservative or even invalid, in which case a seemingly "strict" safety guarantee may be misleading; in such settings we recommend either (i) increasing offline overlap ρ and re-estimating $\underline{R}_0$ with more robust methods, or (ii) choosing a larger β to reduce reliance on a potentially brittle lower bound.

Putting the knobs together. Design is easiest to think of sequentially: first ensure overlap (choose $\rho > 0$ if logging is under our control), then decide how much offline data n is worth given the intended online horizon, then select ε to balance discretization against (offline-subsidized) demand learning, and finally calibrate β to the credibility of offline baseline evaluation and the institution's tolerance for short-run revenue drawdowns. This ordering mirrors the model's message: the algorithm illuminates the tradeoff between experimentation and revenue protection, and offline information shifts that frontier outward by making learning less disruptive.

7 Comparative statics and design guidance

The main design lesson from the offline-to-online analysis is that the "cost" of learning can be summarized by a small set of information objects that practitioners can often influence: (i) how much randomized overlap the logs contain, (ii) how large and how diverse the logged contexts are, and (iii) how finely we discretize the one-dimensional increment axis used to learn demand. We emphasize intuition first: offline randomization makes experimentation *less visible* and often cheaper (it is already "paid for"), and its value is greatest precisely when the online horizon is short or when conservative deployment constraints limit aggressive online probing.

Logged overlap ρ: turning identifiability on, and reducing online disruption. When the logging propensity has a uniform component with weight $\rho > 0$, the log contains rounds in which price variation is exogenous relative to x. This matters twice. First, it delivers unbiased valuation-index signals for estimating θ, effectively yielding an offline regression sample size on the order of ρn. Second, those same rounds enlarge the initial design matrix V_0, which shrinks the geometry-driven uncertainty that triggers valuation-approximation ("exploration") behavior online. A useful way to see diminishing returns is through the log-determinant:

$$\log\det(V_0) = \log\det\Big(I + \sum_{i\in\mathcal{I}_u} x_i^L (x_i^L)^\top\Big) \;\le\; d\log\Big(1 + \tfrac{B_x^2|\mathcal{I}_u|}{d}\Big) \;\approx\; d\log\Big(1 + \tfrac{B_x^2\rho n}{d}\Big).$$

Because the right-hand side grows only logarithmically in ρn, increasing ρ from 0 to a small positive value is often qualitatively transformative (it makes the warm start feasible and reduces the earliest, most uncertain online phase), whereas increasing ρ from moderate to large yields more incremental gains. In operational terms, if we can choose the logging policy, a small amount of persistent randomization can be viewed as an investment that lowers future experimentation intensity in deployment, where random prices may be more salient to users and stakeholders.

Log size n: value comes from coverage, not just volume. Scaling up n improves both sides of the learning problem. On the valuation side, larger n (at fixed ρ) increases $\det(V_0)$ and thus reduces the number of online rounds in which we must "spend" on valuation approximation. On the demand side, more logs raise the bin counts $N_{k,0}$ used to initialize $\hat{D}_{k,0}$, tightening early confidence bands and accelerating elimination of dominated increments. The economically relevant nuance is that *effective* log size is about coverage: if contexts are concentrated in a low-dimensional subspace or if the randomized component occurs primarily in a narrow segment of the price range, then V_0 may be ill-conditioned and many increment bins may remain empty. Hence, when logs are collected intentionally, we recommend prioritizing (i) broad context diversity (to avoid weak directions in V_0) and (ii) nontrivial price support (to populate bins across the increments that are likely to be relevant for optimal pricing).

Grid precision ε: choosing where to pay bias, and where to pay variance. The increment grid resolution ε governs a classic bias–variance tradeoff. Finer grids reduce discretization error in price optimization but require more information per increment to learn $D(\delta)$ with confidence. A convenient way to organize tuning is to compare the magnitudes of the discretization term and the learning-complexity term in the regret decomposi-

tion (suppressing logs and constants):

$$\text{discretization} \sim T\varepsilon, \qquad \text{demand learning} \sim \varepsilon^{-2}.$$

Offline initialization effectively subsidizes the second term: each bin begins with $N_{k,0}$ pseudo-observations, so the algorithm enters the online phase with narrower demand confidence intervals in well-populated regions. This suggests a pragmatic refinement to the theoretical default $\varepsilon = (d^2 \log^2 T/T)^{1/3}$: if the log is large and the randomized prices cover the relevant increment range, we can reduce ε modestly to lower discretization loss without suffering an equivalent increase in online exploration. Conversely, if many bins are empty or near-empty, overly fine grids create a long tail of poorly estimated increments; in that case, we prefer either (i) truncating $\mathcal{K}$ to the empirically supported region, or (ii) using nonuniform bin widths (coarser where $N_{k,0}$ is small), accepting local bias in exchange for stable inference.

Safety slack β: conservatism should track evaluation credibility. When we wrap learning with a baseline-safety budget rule, the parameter β controls how much short-run revenue drawdown we are willing to tolerate relative to the baseline lower bound $\underline{R}_0$. Mechanically, larger β relaxes the constraint and reduces forced baseline plays; smaller β enforces stricter protection but can slow learning. The deeper point is that β should be calibrated to the *trustworthiness* of $\underline{R}_0$, which depends on overlap quality and propensity correctness. If propensity models are misspecified, or if the context distribution shifts between logging and deployment, then an apparently conservative $\underline{R}_0$ may not be a valid floor. In such environments we should either strengthen the evidence behind $\underline{R}_0$ (e.g., increase ρ, improve propensity logging, use more robust evaluation) or choose a larger β to reduce dependence on a potentially brittle certificate.

Actionable tuning checklist. We recommend the following workflow. First, ensure overlap: if logging is under our control, enforce $\rho > 0$ and monitor realized randomized coverage. Second, assess whether V_0 is well-conditioned (or equivalently whether $\log\det(V_0)$ is substantial) and whether bins with relevant δ_k have adequate $N_{k,0}$. Third, start from the default ε and adjust based on empirical bin support, preferring truncation or adaptive binning to extremely fine uniform grids. Finally, set β in light of institutional risk tolerance *and* the reliability of offline evaluation, recognizing that stricter guarantees can be counterproductive if the baseline certificate is itself uncertain. This sequencing reflects the model's central tradeoff: offline information shifts the feasible frontier by making online learning less disruptive, but only to the extent that the logs provide genuine overlap and stable measurement.

8 Simulations and empirical template (optional)

The regret bounds and comparative statics above are most interpretable when we can *see* how the warm start changes the algorithm's early behavior. A minimal simulation template therefore focuses on two observables that the theory highlights: (i) cumulative revenue (or regret) and (ii) the incidence of "disruptive" learning events, i.e., valuation-approximation rounds $\mathcal{G}_{on}$ and (when enabled) forced baseline plays under the conservative wrapper. Our goal in this subsection is not to optimize performance, but to provide an empirical checklist that mirrors the model's primitives and makes the role of logged overlap ρ operational.

Synthetic, drift-free environment. We recommend starting with a stationary instance exactly matching the analysis. Fix $(d, T, B_x, B_y, B_\theta, B_\xi, L_\xi)$ and choose a context distribution $\mathcal{P}_x$ with $\|x_t\|_2 \le B_x$ almost surely (e.g., $x_t \sim \mathrm{Unif}(\{\pm 1\}^d)$ rescaled, or a truncated Gaussian normalized to radius B_x). Draw θ once with $\|\theta\|_2 \le B_\theta$. For the noise, pick a bounded distribution with a Lipschitz CDF, e.g. $\xi_t \sim \mathrm{Unif}([-B_\xi, B_\xi])$ (then $L_\xi = 1/(2B_\xi)$) or a truncated logistic; generate valuations $y_t = x_t^\top \theta + \xi_t$ and binary outcomes $o_t = \mathbf{1}\{p_t \le y_t\}$.

Offline logs are generated by the known mixture propensity

$$\mu_0(p \mid x) = \rho\, \mathrm{Unif}([-B_y, B_y]) + (1-\rho)\, \delta_{\pi_0(x)},$$

so that we can identify $\mathcal{I}_u$ and compute $(V_0, \hat{\theta}_0, \hat{D}_{k,0}, N_{k,0})$. For π_0, we suggest a simple, intentionally imperfect policy to make improvement visible (e.g. a constant price, or a linear heuristic clipped to $[0, B_y]$). Online, run (i) cold-start VAPE and (ii) offline-initialized VAPE with the same $\varepsilon = (d^2 \log^2 T/T)^{1/3}$ and confidence $\alpha = T^{-4}$.

Primary outcome plots: what the warm start changes. We find the following plots most diagnostic:

- *Cumulative regret* proxy: since $\max_p \pi(x_t, p)$ is known in simulation, report $\sum_{t \le s} \left(\max_{p \in [0, B_y]} pD(p - x_t^\top \theta) - p_t D(p_t - x_t^\top \theta) \right)$ as a function of s.
- *Exploration incidence*: plot the running count $\sum_{t \le s} \mathbf{1}\{t \in \mathcal{G}_{on}\}$ for cold vs warm start. This figure directly visualizes Proposition 3: warm-start curves should be uniformly below cold-start curves, with the largest gap early in time.
- *Demand learning coverage*: histogram the offline bin counts $\{N_{k,0}\}_k$ and, optionally, the total counts after T online rounds. This connects performance to whether the increment grid is supported where the algorithm searches.

The qualitative pattern consistent with the theory is that warm starts compress the "uncertain" initial phase: the policy reaches stable pricing sooner, and the reduction in $|\mathcal{G}_{on}|$ is largest when T is modest (when early mistakes are relatively more costly).

Ablations: isolating mechanisms and the role of ρ. To attribute improvements to specific offline objects, we recommend three ablations.

1. *Valuation-only warm start*: initialize $(V_0, \hat{\theta}_0)$ but set $\hat{D}_{k,0}$ uninformative (e.g. $N_{k,0} = 0$). This isolates exploration reduction through $\log\det(V_0)$.

2. *Demand-only warm start*: keep $V_0 = I$ but initialize $(\hat{D}_{k,0}, N_{k,0})$ from logs using an oracle θ (or using $\hat{\theta}_0$ while not feeding V_0 to the online estimator). This isolates whether early demand confidence drives most of the gain.

3. *Vary ρ at fixed n*: run a sweep over $\rho \in \{0, 0.01, 0.05, 0.1, 0.2\}$ (or feasible values in the application). The central empirical prediction is a sharp change between $\rho = 0$ and small $\rho > 0$ (identifiability and warm start "turn on"), followed by diminishing returns as ρ grows, consistent with the logarithmic growth of $\log\det(V_0)$.

Conservative wrapper: safety–learning tradeoffs in a controlled setting. To illustrate the safety mechanism, compute an offline lower confidence bound $\underline{R}_0$ for the baseline value (IPS or DR is straightforward here because μ_0 is known), then run initialized VAPE (a) without the wrapper and (b) with the budget recursion. Report: (i) the realized minimum of the budget process $\min_{s \le T} B_s$, (ii) the number of forced baseline plays, and (iii) realized revenue shortfall relative to $(1-\beta)s\underline{R}_0$ over time. In stationary simulations with correct propensities, we typically see that modest slack β eliminates most forced plays after a short transient, making safety nearly costless once uncertainty contracts; aggressive safety (small β) visibly slows early learning.

Mild-drift stress test: robustness diagnostics rather than guarantees. Finally, because deployment rarely matches the logging distribution exactly, we recommend a "mild drift" variant as a stress test. Two simple perturbations are: (i) parameter drift $\theta_t = \theta + \Delta \cdot \mathbf{1}\{t > t_0\}$ with small $\|\Delta\|_2$, and (ii) context shift where x_t changes distribution after t_0 while preserving $\|x_t\| \le B_x$. These experiments typically show that warm starts still reduce early disruption, but the conservative wrapper's behavior becomes more sensitive to whether $\underline{R}_0$ remains a valid certificate under shift. This motivates the next section: when propensities are unknown, overlap is weak,

or stationarity fails, we generally need additional estimation and robustness machinery beyond the clean offline-to-online template.

9 Limitations and extensions

The offline-to-online template above is intentionally clean: we assume the logging propensities $\mu_0(p \mid x)$ are known, overlap is controlled by an exogenous $\rho > 0$, and the offline and online environments share the same stationary primitives. These assumptions let the model isolate a transparent economic logic—we can "buy" safer and faster online learning with randomized historical experimentation—but they also mark the boundary beyond which additional statistical and numerical machinery becomes essential.

Unknown or misspecified propensities. In practice, platforms often do not have perfectly reliable records of the randomization mechanism that generated historical prices (or the mechanism itself may have evolved). If μ_0 is unknown, then both (i) our ability to identify θ from the uniform component and (ii) the validity of IPS/DR baseline evaluation are threatened. From an econometric perspective, the issue is classical: propensity misspecification induces bias in off-policy evaluation, and a biased "lower confidence bound" $\underline{R}_0$ can invalidate any safety guarantee. A pragmatic extension is to explicitly estimate $\mu_0(p \mid x)$ from logs (e.g., multinomial/continuous density models for $p \mid x$) and then treat propensity estimation error as part of the confidence radius. Doing so typically requires sample splitting or cross-fitting to avoid overly optimistic concentration when the same data are used to fit and evaluate propensities. When the price distribution has a discrete atom at $\pi_0(x)$ plus a continuous component, the numerical implementation is also delicate: one must fit a mixture model and enforce positivity constraints to prevent exploding weights. In deployments, we view instrumentation (logging the randomization seed and assignment probabilities) as a first-order policy recommendation, because it substitutes an organizational control for a statistical correction.

Weak overlap and partial identification. Our warm-start gains scale with the "effective randomized sample size" $|\mathcal{I}_u| \approx \rho n$. When ρ is very small, $\det(V_0)$ may barely exceed 1, demand bins $\{N_{k,0}\}_k$ become sparse, and the offline stage cannot materially reduce disruptive online exploration. More fundamentally, if the logged prices concentrate on a narrow range, then counterfactual revenues outside that range are not point-identified without functional-form restrictions. In that regime, one typically moves from regret analysis under point identification to *robust* or *set-identified* objectives: we can optimize a pessimistic revenue criterion over a plausible model class, or

impose a conservative constraint that only certifies improvements where overlap exists. Algorithmically, this leads to practices such as weight clipping, pessimistic value iteration over confidence sets, or explicitly allocating *new* randomized exploration online (a small "tax" on short-run revenue to restore identifiability). Economically, the message is that a deterministic incumbent policy π_0 without experimentation creates an informational externality: it preserves short-run stability at the cost of learning opportunities that would enable future surplus.

Nonstationary logs versus online environments. A second fault line is distribution shift: the log-generating environment may differ from the online environment due to seasonality, product changes, market entry, or simply a different context distribution $\mathcal{P}_x$. Our regret analysis tolerates stochastic contexts, but it does not protect against a structural break between the offline and online regimes. The most direct consequence is that $\underline{R}_0$, even if valid offline, may no longer be a valid certificate online, so the conservative wrapper can become either overly restrictive (if the baseline improves) or unsafe (if the baseline deteriorates). Extensions here resemble the toolkit for covariate shift and nonstationarity: one can (i) reweight offline observations by an estimated density ratio between online and offline contexts, (ii) maintain a time-varying baseline certificate updated with online data, or (iii) replace static regret with dynamic regret benchmarks that allow θ_t to drift. Each approach introduces nontrivial numerical steps (density-ratio estimation, online change-point detection, or sliding-window estimation) and typically weakens the clean $\tilde{O}(T^{2/3})$ guarantee into bounds that depend on a variation budget such as $\sum_t \|\theta_{t+1} - \theta_t\|_2$.

Context-dependent noise and richer demand heterogeneity. Our model assumes ξ_t is i.i.d. with a common Lipschitz CDF F, so the demand increment function $D(\delta) = \mathbb{P}(\xi \geq \delta)$ is context-invariant. In many markets, however, uncertainty is heteroskedastic or systematically linked to x_t (e.g., different customer segments have different dispersion). Formally, this corresponds to $F(\cdot \mid x)$ and hence $D_x(\delta)$. Once D becomes context-dependent, the increment-binning step must be redesigned: the same δ_k no longer aggregates comparable observations across contexts, and the offline histogram is no longer estimating a single object. One extension is to posit a structured family, such as a scale model $\xi = \sigma(x)u$ or a generalized linear specification for purchase probabilities, and to estimate nuisance functions $\sigma(\cdot)$ or a link function using flexible methods. This is precisely where doubly robust techniques become attractive: we can combine an estimated outcome model (purchase probability) with propensity estimates to stabilize learning. The price of flexibility is computational: fitting high-dimensional nuisance models, tuning regularization, and propagating their uncertainty into valid

confidence sets generally requires cross-fitting and careful finite-sample calibration.

Takeaway. We see the present framework as a baseline: it illuminates the safety–learning tradeoff when logged experimentation is available and correctly recorded. Moving from this benchmark to production requires confronting propensity uncertainty, overlap scarcity, and environmental drift, and doing so typically shifts the bottleneck from analytic regret decomposition to reliable numerical estimation of nuisance components and robust certification under misspecification.