Why Offline Pricing RL Fails: Non-Identification under Confounding and a Minimal-Randomization Fix with Finite-Sample Bounds

Liz Lemma Future Detective

January 16, 2026

Abstract

Dynamic retail pricing work (including Q-learning approaches such as the source paper) commonly evaluates policies in simulated environments or from historical logs as if demand were conditionally identified given observed features. In modern 2026 retail stacks, price is chosen jointly with marketing, ranking, inventory throttles, and targeting, creating hidden confounding; logs also exhibit limited price support because production systems avoid 'bad' prices. This paper formalizes a clean negative result: without overlap or a valid source of randomization, offline evaluation of a new pricing policy is not point-identified, and no estimator can guarantee accurate policy-value estimation. We then provide a constructive remedy that is operationally minimal: introduce a small, carefully designed randomized price perturbation on a subset of traffic, which acts as an instrument. Under tractable assumptions, this restores identifiability and yields tight, implementable finite-sample confidence intervals for policy value (and for constraint metrics such as groupwise price fairness). We give partial-identification bounds under bounded-confounding sensitivity models, minimax lower bounds showing when failure is unavoidable, and an exploration design that optimizes information gain subject to a revenue-loss cap. Empirical illustrations use semi-synthetic retail environments with injected confounders to mimic production pricing systems.

Table of Contents

3. 3. Baseline model and estimands: logged data, target policy value, and constraint metrics (stability/fairness) as functionals of counterfactual demand.

4. 4. Non-identification results: sharp examples (binary prices) showing lack of overlap and hidden confounding imply a wide identified set; minimax lower bounds for any estimator.

5. 5. Partial identification under bounded confounding: sensitivity model (e.g., odds-ratio bound Γ or additive bias bound) and closed-form bounds on policy value in a tractable binary-price setting; extensions to multi-price via linear programming bounds.

6. 6. Minimal randomization as an instrument: design of a small exploration policy with probability ρ, local perturbation constraints, and identifiability under instrument exogeneity; discussion of operational feasibility.

7. 7. Estimation and inference: doubly robust / IV-style estimators exploiting known randomization propensities; finite-sample concentration and confidence intervals; guidance on choosing ρ to meet a revenue-loss cap.

8. 8. Semi-synthetic experiments: inject confounders (marketing shocks, rank changes, stockouts), compare naive OPE, bounded-confounding bounds, and the minimal-randomization estimator; sensitivity to ρ, overlap, and policy shift magnitude.

9. 9. Implications for RL pricing: what can/can't be learned from logs; recommended instrumentation standards for safe deployment; extensions to multi-SKU and competitive settings.

10. 10. Conclusion and policy/audit checklist: minimal requirements for credible claims (profit lift, fairness), limitations, and future work.

1 Introduction

Dynamic pricing systems sit at an uneasy intersection of operations research, econometrics, and modern machine learning. On the one hand, retailers now have the engineering capacity to deploy rich contextual policies that map high-dimensional signals—traffic, competitor prices, inventory risk, user segments, and channel conditions—into a posted price in near real time. On the other hand, the economic object we ultimately care about is not predictive accuracy but counterfactual profit: what would revenue and margin have been had we followed a new pricing rule rather than the incumbent one? This gap between what we can optimize in silico and what we can justify in deployment has become the central bottleneck for dynamic pricing in 2026, especially as firms attempt to move from carefully tuned rule-based price ladders to bandit and reinforcement learning (RL) pipelines that adapt endogenously to market feedback.

A practical reason for the bottleneck is that offline logs are not experimental data. The historical prices we observe are typically produced by a production stack that reacts to information not recorded in the evaluation table: merchandising pushes, search ranking boosts, inventory throttling, quality-of-traffic shocks, and latent demand news that a human merchant sees before it reaches the feature store. Even when analysts control for a rich X, the realized price may still load on an unobserved state that simultaneously shifts demand. In such environments, a naive regression of demand on price, or a standard inverse-propensity reweighting that treats the logging propensities as functions of observables alone, can deliver precisely the wrong lesson: the algorithm appears to "learn" a demand curve that is partly a story about selection. The consequence is not a small-sample nuisance but a conceptual one: there need not exist a unique mapping from the joint distribution of logged observables to the policy value of a counterfactual pricing rule.

A second, and equally pervasive, obstacle is limited support. Modern pricing stacks implement guardrails for brand and margin, enforce discrete ladders, and often couple price changes to calendar and inventory regimes. As a result, conditional on a given context, many prices are never shown. This lack of overlap is particularly acute for the policies we most want to evaluate: policies that would deliberately shift mass to rarely used prices (e.g., to test higher prices on low-elasticity segments) or that would smooth prices across regimes where the incumbent system behaves discontinuously. From the standpoint of offline evaluation, this creates a knife-edge problem: a counterfactual policy can place positive probability on actions that have zero empirical support in precisely those contexts that matter for profit. Without additional structure, the data cannot speak about those counterfactuals, and any algorithm that reports a single "best estimate" of profit is implicitly making untestable extrapolations.

These two features—hidden confounding and limited support—explain why the celebrated promise of RL for pricing has been slower to materialize than the algorithmic literature might suggest. RL is well suited to sequential decision problems when exploration is possible and reward feedback is informative. Yet the deployment reality of pricing is that exploration is costly, and most organizations insist on strong guardrails precisely because early mistakes are visible in revenue. This produces a paradox: the less the firm is willing to explore, the less it can reliably evaluate or improve its policy; but the less it can evaluate, the harder it is to justify exploration. Our goal is to make this tradeoff explicit, and to provide a set of tools that are honest about what can be learned from logs, while still offering an actionable path toward credible evaluation with minimal disruption.

We make three contributions. First, we formalize the sense in which off-policy evaluation in pricing can fail sharply. In a stylized but economically meaningful setting, we show that when the incumbent system selects prices based on an unobserved state and fails to provide overlap, the value of a target policy is not point-identified: multiple demand worlds fit the same log distribution yet imply very different profits under the counterfactual policy. This non-identification is not merely philosophical; it yields minimax lower bounds indicating that no estimator can guarantee small error uniformly over plausible environments. The implication for practice is immediate: "offline A/B tests" computed from observational logs can be arbitrarily misleading, even with massive datasets, when the missing variation is structural rather than statistical.

Second, we develop a partial-identification approach that replaces point estimates with economically interpretable bounds. Rather than assuming away the unobserved state, we allow it but restrict how strongly it can tilt price selection through a sensitivity parameter that bounds the degree of confounding. This delivers an identified interval for the counterfactual value, trading sharpness for transparency. In settings where leadership is unwilling to randomize prices broadly, such robustness analysis provides a disciplined way to ask: how large would hidden selection need to be to overturn the apparent profitability of a proposed policy? We emphasize that these bounds are not a substitute for experimentation; they are a diagnostic that quantifies how much of the conclusion is driven by unverifiable assumptions.

Third, and most importantly for deployment, we show how a small amount of designed randomization can restore identifiability. We study an "exploration slice" in which a fraction of traffic is assigned prices according to a known, context-dependent randomization distribution. This is conceptually modest—it need not replace the incumbent system, and it can respect operational guardrails through careful choice of support—but it is economically powerful because it breaks the link between price and the unobserved state on the randomized slice. We then provide an implementable estimator that uses this slice to recover causal demand and yields finite-sample confidence intervals whose width scales with the effective randomized sample size. The operational message is that credible offline evaluation is not a purely statistical problem; it is a systems-design problem, and a well-engineered instrument can convert an intractable identification problem into a standard one.

Our analysis also connects to emerging concerns beyond average profit. Pricing policies are increasingly audited for distributional impacts across groups and channels, and regulators as well as internal risk teams ask for evidence that algorithmic price changes do not induce unjustified disparities. The same obstacles that hinder profit evaluation—confounding and lack of overlap—also hinder credible auditing. A virtue of the minimal-randomization approach is that it produces a clean basis for both value estimation and policy diagnostics on the same experimental slice, albeit at the cost of deliberate exploration.

The paper proceeds as follows. Section 2 situates our contribution within the dynamic pricing, bandit/RL, off-policy evaluation, and causal inference literatures. We then introduce the model primitives and the offline evaluation target, highlighting where production systems generate hidden confounding and support restrictions. Next, we establish sharp non-identification and minimax impossibility results that clarify the limits of purely observational evaluation. We then present bounded-confounding identification regions and discuss their interpretation as sensitivity analyses. Finally, we turn to the design and analysis of minimal randomization, deriving identification, proposing estimators, and providing finite-sample inference guarantees. We close by discussing implementation considerations—guardrails, exploration budgeting, and monitoring—and by acknowledging limitations, including the assumptions required for the instrument to be valid and the organizational constraints that shape feasible experimentation.

2 Related Literature

Our setting sits at the boundary of several literatures that have largely progressed in parallel: (i) classical dynamic pricing in operations research and revenue management, (ii) learning-based pricing via bandits and reinforcement learning, (iii) off-policy evaluation and causal inference with observational logs, (iv) partial identification and sensitivity analysis under hidden confounding, and (v) algorithmic auditing, including distributional and fairness diagnostics for deployed decision rules. We briefly position our contribution relative to each.

The revenue-management tradition studies pricing as a control problem under a demand model, often with explicit inventory, capacity, and time dynamics **??**. A central methodological move in this literature is to impose structure—parametric demand curves, monotonicity, concavity, or bounded elasticities—that turns pricing into a tractable optimization problem and yields policy insights (e.g., bid prices and protection levels). In modern marketplace implementations, these models are frequently operationalized through context-enriched demand forecasts and guardrails (price ladders, brand constraints, and margin rules). Our focus is complementary: we take the production reality of complex contextual pricing as given and ask what can be inferred about the profit of a *counterfactual* policy from logs that are neither randomized nor guaranteed to exhibit overlap. This perspective is closer in spirit to econometrics than to classical RM optimization: the primary bottleneck is not solving the retailer's dynamic program conditional on a known demand system, but rather identifying the relevant demand object in the first place when the historical policy responds to information not captured in the evaluation table.

A second strand emphasizes *learning* demand while simultaneously pricing. Multi-armed bandit and reinforcement-learning formulations treat prices as actions and sales as rewards, aiming for low regret relative to an optimal policy **???**. The technical success of this literature relies on exploration: the algorithm deliberately perturbs prices to learn the demand response, and regret analyses quantify the cost of experimentation. In contrast, the empirical deployments we have in mind typically begin with an incumbent policy that was never designed as an exploration mechanism, and the organization often seeks "offline validation" before permitting meaningful experimentation. This institutional sequencing reverses the usual bandit logic and creates the paradox highlighted in the introduction: without exploration, offline evaluation is fragile; yet without credible evaluation, exploration is hard to justify. Our minimal-randomization instrument can be read as a practical bridge between these worlds: it injects a small, auditable amount of exploration into an otherwise production-grade system, making causal learning and evaluation feasible while preserving operational guardrails.

The methodological core of our paper connects most directly to the off-policy evaluation (OPE) literature in machine learning and to semiparametric causal inference. When propensities are known (or estimable) and overlap holds, inverse propensity scoring, direct regression, and doubly robust estimators provide consistent value estimates of a target policy **??**. Recent work strengthens these tools through cross-fitting, orthogonalization, and finite-sample concentration for bounded rewards, yielding practical estimators with uncertainty quantification **??**. Our contribution highlights a failure mode that is especially salient in pricing: the historical policy may depend on unobservables U that also shift demand, so the "propensity" relevant for identification is $\mu_0(p \mid x, u)$, not $\mu_0(p \mid x)$. In that case, standard OPE estimators can be inconsistent even with infinite data, and when support is limited the problem can be worse: there may be no data at all for the actions the target policy would choose in important contexts. We formalize these points as non-identification and minimax impossibility results, not as critiques of OPE per se, but as a reminder that OPE is ultimately a *design* problem: identification requires either credible ignorability or an instrument that restores it.

A natural response to hidden confounding is to relax point identification and instead report *bounds*. This approach has deep roots in econometrics through partial identification **?** and in causal inference through sensitivity analysis **?**. In pricing applications, bounds are appealing because they translate unverifiable assumptions into explicit economic tradeoffs: one asks how strong selection on unobservables must be to overturn a profitability claim. Our bounded-confounding analysis follows this logic by indexing the identified set for $V(\pi)$ with a sensitivity parameter Γ that limits how much unobserved states can tilt assignment odds. Related ideas appear in robust policy learning and distributionally robust OPE, where uncertainty sets are placed on propensities or outcome models to obtain worst-case guarantees **?**. We view such robustness tools as complementary to experimentation: they are most useful as diagnostics when leadership constrains exploration, and as a way to communicate the fragility of purely observational conclusions.

Our emphasis on a small randomized “exploration slice” also connects to the econometric literature on instrumental variables and encouragement designs, as well as to modern “interleaving” and switchback experiments in marketplaces. In many online systems, fully randomized experimentation is infeasible because treatments interact through congestion, ranking, or inventory; nonetheless, carefully scoped randomization can identify local causal effects under exclusion restrictions and stability assumptions **?**. In pricing, the relevant exclusion restriction is that the instrument Z affects demand only through the posted price, and the key operational requirement is that the randomized prices respect guardrails so that the experiment is ethically and commercially acceptable. Our model abstracts from interference across units, but the deployment message aligns with marketplace practice: even limited, well-instrumented randomization can be more informative than massive observational logs when the missing variation is structural.

Finally, our discussion of constraint metrics, including group-level price and profit disparities, relates to a growing literature on algorithmic auditing and fairness in decision systems **??**. Most formal fairness criteria were developed for classification and risk scoring, but pricing raises distinct issues: the action is continuous (or high-cardinality discrete), payoffs are monetary, and protected attributes may enter both demand and cost. Moreover, the same confounding and overlap problems that plague value estimation also plague auditing: if certain prices are never shown to certain groups (or only shown under unobserved promotional states), disparity estimates computed from logs can be artifacts of selection. Our framework treats these diagnostics as functionals of counterfactual demand under a policy, making clear when they are identified, when they are only partially identified, and when minimal randomization can ground auditing in experimental variation.

Taken together, the literatures above motivate our central thesis: credible evaluation of pricing policies is not guaranteed by scale or sophistication of the learning algorithm; it hinges on identification, which in turn hinges on overlap and on the relationship between the deployed policy and unobserved drivers of demand. The next section formalizes the baseline model, the value functional $V(\pi)$, and the additional constraint metrics that we use to evaluate policies beyond average profit.

3 Baseline Model and Estimands

We evaluate pricing policies in a setting where the retailer observes rich contextual signals but only a narrow and potentially selection-distorted slice of the price–demand relationship in historical logs. The unit of analysis is a single pricing "interaction" (e.g., a page view, session, or decision window for a SKU), indexed by t online and by $i = 1, \ldots, n$ in offline data. In each interaction the environment draws observable context X (product attributes, calendar effects, traffic conditions, user segment, channel) and an unobserved state U that captures latent demand shocks and operational factors that are rarely logged cleanly (marketing intensity, ranking boosts, throttling, or inventory pressure). The retailer posts a price $P \in \mathcal{P}$ and then demand Y realizes (units sold, or a purchase indicator). We also observe unit cost C, which may depend on X and time.

Formally, the historical dataset is

$$\mathcal{D}_n = \{(X_i, P_i, Y_i, C_i)\}_{i=1}^n,$$

generated by a logging policy μ_0 that may depend on both observed and unobserved states:

$$(X_i, U_i) \sim \mathbb{P}, \qquad P_i \sim \mu_0(\cdot \mid X_i, U_i), \qquad Y_i \sim \mathbb{P}(\cdot \mid X_i, P_i, U_i).$$

The key feature is that U can simultaneously influence the chosen price and realized demand, so that $P \not\perp U \mid X$ and $Y \not\perp U \mid (X, P)$. This is not a modeling nuisance; it is a faithful abstraction of production systems in which pricing responds to information outside the evaluation table. Importantly, our baseline model does *not* assume overlap: for some contexts x, the realized support of $P \mid X = x$ can be a strict subset of $\mathcal{P}$ (e.g., a price ladder where only a few rungs were ever deployed for a given segment).

To define what we want to evaluate, we adopt the potential-outcomes representation. For each interaction and each feasible price $p \in \mathcal{P}$, let $Y(p)$ denote the demand that would realize if price p were posted. We impose the usual consistency requirement, $Y = Y(P)$, which states that the observed outcome equals the potential outcome under the posted price. We do not require any form of ignorability; in fact, the possibility that $Y(p)$ is correlated with the realized price via U is the central empirical difficulty. Nonetheless, the counterfactual object $Y(p)$ provides a clean language for estimands: it separates what a policy would do (choose prices) from how the environment would respond (generate demand).

A (possibly stochastic) target pricing policy is a conditional distribution $\pi(p \mid x)$ over $\mathcal{P}$. Stochasticity is useful both conceptually (it nests randomized or softened policies) and practically (modern systems often sample among a set of candidate prices). Given π, we define per-interaction profit as

$$R(X, P, Y, C) = (P - C)\, Y,$$

and the target policy value (expected profit) as the counterfactual functional

$$\begin{aligned} V(\pi) &= \mathbb{E}\big[(P-C)Y \,\big|\, P \sim \pi(\cdot \mid X)\big] && (1) \\ &= \mathbb{E}\left[\sum_{p \in \mathcal{P}} \pi(p \mid X)\,(p - C)\,Y(p)\right], \qquad \text{(discrete } \mathcal{P}\text{)}. && (2) \end{aligned}$$

When $\mathcal{P}$ is continuous, the summation is replaced by an integral. Equation (2) highlights the two primitive ingredients needed to evaluate a counterfactual pricing rule: the policy itself, and the conditional causal demand curve $m(p, x) := \mathbb{E}[Y(p) \mid X = x]$ (together with the cost process). In many applications one may posit a structural or working demand model $D(p, x)$ (e.g., log-linear or isoelastic) to regularize estimation; we treat such structure as optional and, when used, as a maintained assumption whose credibility must be judged in light of the support and confounding issues above.

Beyond average profit, organizations typically impose constraints and diagnostic metrics that are also counterfactual functionals of the same objects. We group these into two broad classes. The first are *stability and operational* metrics, motivated by guardrails that prevent disruptive price paths or excessive dispersion. Examples include the average posted price under the target policy,

$$\mathbb{E}_\pi[P] = \mathbb{E}\left[\sum_{p \in \mathcal{P}} \pi(p \mid X)\, p\right],$$

and context-conditional dispersion (useful for detecting policies that randomize too aggressively in sensitive regions),

$$\mathbb{E}[\mathrm{Var}_\pi(P \mid X)] = \mathbb{E}\left[\sum_p \pi(p \mid X)\, p^2 - \Big(\sum_p \pi(p \mid X)\, p\Big)^2\right].$$

In settings where interactions are temporally linked (e.g., a SKU-day panel), one may also care about intertemporal stability such as $\mathbb{E}[|P_t - P_{t-1}|]$; our analysis is written at the interaction level, but these pathwise metrics can be handled by defining X to include lagged state and by treating the resulting functionals as part of the evaluation target.

The second class are *distributional and fairness* metrics, which ask how a policy allocates prices and profits across groups. Let $G \in \{0, 1\}$ be a group label included in X (e.g., region, platform, or a protected attribute when legally and ethically appropriate). Two simple diagnostics are a price disparity and a profit disparity:

$$\Delta_{\text{price}}(\pi) = \mathbb{E}[P \mid G = 1, \pi] - \mathbb{E}[P \mid G = 0, \pi], \qquad \Delta_{\text{profit}}(\pi) = V(\pi \mid G = 1) - V(\pi \mid G = 0),$$

where $V(\pi \mid G = g)$ is defined as in (1) but with the outer expectation taken conditional on $G = g$. More granular metrics can compare quantiles of the induced price distribution, or impose constraints such as $\mathbb{E}[P \mid G = 1, \pi] \leq \mathbb{E}[P \mid G = 0, \pi] + \kappa$ for a tolerance κ. The common structure is that these are all functionals of $(\pi, m(\cdot, \cdot))$ and therefore inherit whatever identification (or non-identification) properties hold for the causal demand curve.

This section's role is to separate *what* we want to know from *what* the logs can reveal. The estimands above are well-defined under minimal causal consistency, even when the observed data are generated by a confounded and support-limited policy μ_0. The next step is to confront the gap between definition and identification: when the historical system chooses prices using unobserved information and fails to explore parts of $\mathcal{P}$, the mapping from the distribution of observables (X, P, Y, C) to the counterfactual objects $m(p, x)$—and hence to $V(\pi)$ and the associated constraint metrics—can be fundamentally many-to-one. The following section makes this failure precise through sharp examples and minimax lower bounds.

4 Non-identification and minimax impossibility

The estimands in Section 3 are meaningful without strong assumptions, but they need not be *learnable* from logged data. The core difficulty is that the mapping from the distribution of observables (X, P, Y, C) to the counterfactual demand curve $m(p, x) = \mathbb{E}[Y(p) \mid X = x]$ can be many-to-one when (i) the logging system does not explore the prices a target policy would use (lack of overlap) and/or (ii) the price is chosen using unobserved signals that also shift demand (hidden confounding). In this section we make this precise through sharp binary-price examples and a minimax lower bound. The point is not that every production system is adversarial, but that without additional design or structure, the logs alone cannot rule out adversarially different counterfactual worlds.

4.1 A sharp binary-price example: confounding can destroy point identification even when both prices are observed

To isolate the role of unobserved confounding, consider $\mathcal{P} = \{p_L, p_H\}$ with $p_H > p_L$, suppress cost (or treat C as fixed), and fix a context value $X = x$.[1] Suppose there is a binary unobserved state $U \in \{0, 1\}$ that the production system observes and uses for pricing, but the evaluator does not observe. A stylized but empirically plausible situation is that U represents a latent demand/operational signal (e.g., a marketing surge, a ranking boost, or inventory throttling) that simultaneously affects both (a) the chosen price and (b) the propensity to buy.

Assume that the logging system is *fully* driven by U:

$$P = \begin{cases} p_H, & U = 1, \\ p_L, & U = 0, \end{cases} \qquad \text{(almost surely given } X = x\text{)}.$$

Thus, the marginal log contains both prices whenever $\mathbb{P}(U = 1 \mid X = x) \in (0, 1)$, but the realized support is *stratified*: the log never shows p_H in the $U = 0$ state and never shows p_L in the $U = 1$ state. Now place only weak restrictions on potential outcomes: for each $p \in \{p_L, p_H\}$, $Y(p) \in [0, \bar{y}]$, and we maintain consistency $Y = Y(P)$. We do *not* assume ignorability, so $Y(p)$ may depend on U.

Consider any target policy π that randomizes between the two prices with positive probability at x, i.e., $\pi(p_L \mid x) \in (0, 1)$. Its value at x depends on the mixture

$$\mathbb{E}[(P - C)Y \mid X = x,\, P \sim \pi(\cdot \mid x)] = \sum_{p \in \{p_L, p_H\}} \pi(p \mid x)\,(p - C)\,\mathbb{E}[Y(p) \mid X = x].$$

The log identifies $\mathbb{E}[Y \mid X = x, P = p_L]$ and $\mathbb{E}[Y \mid X = x, P = p_H]$, but under the assignment rule above these are

$$\mathbb{E}[Y \mid X = x, P = p_L] = \mathbb{E}[Y(p_L) \mid X = x, U = 0], \qquad \mathbb{E}[Y \mid X = x, P = p_H] = \mathbb{E}[Y(p_H) \mid X = x, U =$$

Crucially, the counterfactual terms $\mathbb{E}[Y(p_H) \mid X = x, U = 0]$ and $\mathbb{E}[Y(p_L) \mid X = x, U = 1]$ never appear in the observables, because those (U, P) pairs never occur. Consequently, many distinct causal demand models induce the *same* distribution of (P, Y) in the log but imply different values under π.

We can make the non-identification sharp by constructing two data-generating processes that agree on the joint distribution of observables but disagree on the missing counterfactuals. Fix any distribution of (U, Y) under the realized pairs $(U = 0, P = p_L)$ and $(U = 1, P = p_H)$ so that the observed conditional means match the log. Then define two worlds:

[1] Allowing X to vary only strengthens the conclusion: one can apply the argument pointwise in x and then average over X.

- World A sets the unobserved counterfactuals to be low: $Y(p_H) = 0$ when $U = 0$, and $Y(p_L) = 0$ when $U = 1$.
- World B sets them to be high: $Y(p_H) = \bar{y}$ when $U = 0$, and $Y(p_L) = \bar{y}$ when $U = 1$.

Both worlds reproduce the same (X, P, Y) distribution under the logging policy, because the only potential outcomes that affect observed Y are $Y(p_L)$ for $U = 0$ and $Y(p_H)$ for $U = 1$. Yet, under a policy π that sometimes posts the "off-stratum" price, the expected profit differs. In particular, whenever π posts p_H in the $U = 0$ stratum (which happens with probability $\pi(p_H \mid x)\,\mathbb{P}(U = 0 \mid X = x)$), the two worlds disagree by as much as $(p_H - p_L)\bar{y}$ per interaction up to cost normalization. Aggregating, the identified set for $V(\pi)$ contains an interval whose length is bounded below by a term proportional to the mass of the unobserved stratum and the price gap, e.g.,

$$\text{length}(\mathcal{I}(\pi)) \;\gtrsim\; \mathbb{P}(U = 1 \mid X = x)\,(p_H - p_L)\,\bar{y},$$

and in the extreme can be as wide as $(p_H - p_L)\bar{y}$.[2] Intuitively, without data that mix prices within the latent state, we cannot disentangle whether high observed demand at p_H is due to the price or due to $U = 1$.

Two remarks help interpret this example. First, the failure is not driven by limited sample size: even with infinite data, the missing counterfactuals remain unlearned because they are never revealed. Second, note that the log exhibits *marginal* support for both prices, so a naive overlap check based on $\mathbb{P}(P = p) > 0$ would pass; what fails is overlap in the relevant causal sense, namely that within the latent confounding strata the system does not vary price.

4.2 Minimax lower bounds: without overlap, no estimator can be uniformly accurate

We next formalize an even starker impossibility: if the target policy assigns positive probability to prices that are never observed in some contexts, then no estimator—regardless of functional form, machine learning sophistication, or clever reweighting—can guarantee small error uniformly over a reasonable model class.

Suppose there exists a set of contexts $\mathcal{X}_0$ with $\mathbb{P}(X \in \mathcal{X}_0) > 0$ and a price $p^\star$ such that

$$\pi(p^\star \mid x) > 0 \quad \text{but} \quad \mathbb{P}(P = p^\star \mid X = x) = 0 \qquad \forall x \in \mathcal{X}_0.$$

Then the log contains *no* information about the demand response at $p^\star$ on $\mathcal{X}_0$. Consider two environments $\mathbb{P}_1$ and $\mathbb{P}_2$ that coincide on the full distribution

[2] The exact expression depends on whether the target policy places mass on both prices at x and on which stratum is "missing" for which price, but the economic content is invariant: the gap scales with the price difference and the range of feasible demand.

of observables (X, P, Y, C) under the logging policy (hence are statistically indistinguishable from $\mathcal{D}_n$), but differ in the counterfactual mean $\mathbb{E}[Y(p^\star) \mid X = x]$ on $\mathcal{X}_0$ by a fixed amount. Because π puts positive mass on $p^\star$ in those contexts, the policy values $V_1(\pi)$ and $V_2(\pi)$ differ by some $\epsilon > 0$, while no estimator can tell which world generated the data.

This yields a standard minimax conclusion: there exists $\epsilon > 0$ such that for any estimator $\widehat{V}$ based on $\mathcal{D}_n$,

$$\inf_{\widehat{V}} \sup_{\mathbb{P} \in \mathcal{M}} \mathbb{P}\Big(|\widehat{V} - V(\pi)| \geq \epsilon\Big) \;\geq\; \frac{1}{4},$$

for an appropriate model class $\mathcal{M}$ that permits arbitrary (but bounded) counterfactual outcomes off support. The proof is a two-point argument (Le Cam): since $\mathbb{P}_1$ and $\mathbb{P}_2$ induce the same distribution over the data, any estimator must incur nontrivial error on at least one of them. From a practitioner's perspective, this is a “no free lunch” theorem for off-policy pricing evaluation: if the historical system never tried the prices a new policy wants to use in certain segments, then the profit consequences in those segments are not merely hard to estimate—they are not determined by the log at all.

Taken together, the binary confounding example and the overlap-based minimax bound clarify the tradeoff we face. Without either credible structure on demand or deliberate exploration, logs can be consistent with sharply different causal price–demand relationships. The natural response is not to abandon offline evaluation, but to complement it with either (i) transparent sensitivity models that quantify how much confounding would be required to overturn conclusions, or (ii) small, carefully designed randomization that restores identification. The next section develops the former in a tractable way.

4.3 Partial identification under bounded confounding: a sensitivity model

A pragmatic middle ground between (i) assuming away hidden confounding and (ii) giving up on learning from logs is to ask how large the confounding would need to be to change our conclusions. We formalize this through a *sensitivity model* that restricts the extent to which an unobserved state U can tilt the logged price assignment, and then compute the induced *identified set* for the policy value $V(\pi)$. When the sensitivity restriction is tight we obtain informative bounds; when it is loose the bounds revert to the impossibility discussed above.

Sensitivity parameter and feasible reweightings. For expository clarity we present the binary-price case $\mathcal{P} = \{p_L, p_H\}$ and write $T = \mathbf{1}\{P = p_H\}$. Let the (observable) marginal propensity be $e(x) = \mathbb{P}(T = 1 \mid X = x)$. Hidden confounding means that the *true* propensity may depend on U, i.e.,

$e(x,u) = \mathbb{P}(T = 1 \mid X = x, U = u)$. A standard restriction is an odds-ratio bound (Rosenbaum-type):

$$\frac{e(x,u)/(1-e(x,u))}{e(x,u')/(1-e(x,u'))} \in [\Gamma^{-1}, \Gamma] \qquad \forall x,\ \forall u, u', \tag{3}$$

for some $\Gamma \geq 1$. Intuitively, $\Gamma = 1$ corresponds to no hidden confounding (conditional randomization given X), while larger Γ permits the logging system to use U more aggressively in setting prices.

A convenient implication of such bounds is that they restrict the range of *importance weights* that would debias the treated or control slice if $e(x,u)$ were observed. In particular, define the treated weight

$$W_1(x,u) = \frac{e(x)}{e(x,u)} \quad \text{and} \quad W_0(x,u) = \frac{1-e(x)}{1-e(x,u)}.$$

Under (3), both $W_1(x,u)$ and $W_0(x,u)$ are bounded within multiplicative factors controlled by Γ (the exact bounds depend on $e(x)$, but are explicit). Moreover, these weights satisfy a normalization moment:

$$\mathbb{E}[W_1(X,U) \mid X = x, T = 1] = 1, \qquad \mathbb{E}[W_0(X,U) \mid X = x, T = 0] = 1,$$

which expresses that they are Radon–Nikodym derivatives of a feasible reweighting rather than arbitrary scalars.

This motivates the following partial-identification program: for each price p and context x, we allow the conditional distribution of outcomes observed under $P = p$ to be *tilted* by a weight function $w(y)$ that (i) lies in a Γ-dependent interval and (ii) integrates to one. The upper and lower bounds on the counterfactual mean demand $m(p,x) = \mathbb{E}[Y(p) \mid X = x]$ are then the solutions to

$$\overline{m}_\Gamma(p,x) = \sup_w \mathbb{E}[w(Y)\,Y \mid X = x, P = p], \qquad \underline{m}_\Gamma(p,x) = \inf_w \mathbb{E}[w(Y)\,Y \mid X = x, P = p], \tag{4}$$

subject to $w(Y) \in [\underline{w}_\Gamma(x,p), \overline{w}_\Gamma(x,p)]$ almost surely and $\mathbb{E}[w(Y) \mid X = x, P = p] = 1$. The bounds $\underline{w}_\Gamma, \overline{w}_\Gamma$ are directly induced by (3) (or, more conservatively, can be set to $[1/\Gamma, \Gamma]$ as a marginal sensitivity envelope).

Closed-form bounds in the binary-price case. In the binary setting, (4) admits a simple closed form because the objective is linear and the feasible set is a box intersected with a single mean constraint. Since Y is bounded, the extremum is achieved by assigning the *largest* feasible weights to the *largest* outcomes (to maximize) and vice versa (to minimize). Concretely, write $\overline{w} = \overline{w}_\Gamma(x,p)$ and $\underline{w} = \underline{w}_\Gamma(x,p)$. To maximize $\mathbb{E}[w(Y)Y]$ subject to $\underline{w} \leq w(Y) \leq \overline{w}$ and $\mathbb{E}[w(Y)] = 1$, we place weight $\overline{w}$ on an upper tail of

Y and weight $\underline{w}$ on the remainder, with the tail mass chosen to satisfy the mean constraint. Let $\alpha_\Gamma(x,p) \in (0,1)$ solve

$$\alpha_\Gamma(x,p)\,\overline{w} \;+\; (1-\alpha_\Gamma(x,p))\,\underline{w} \;=\; 1,$$

and let $q_\Gamma(x,p)$ be a corresponding $(1-\alpha_\Gamma(x,p))$-quantile of Y under $(X = x, P = p)$. Then one sharp representation is

$$\overline{m}_\Gamma(p,x) = \overline{w}\;\mathbb{E}[Y\,\mathbf{1}\{Y \geq q_\Gamma(x,p)\} \mid X = x, P = p] \;+\; \underline{w}\;\mathbb{E}[Y\,\mathbf{1}\{Y < q_\Gamma(x,p)\} \mid X = x, P = p]\,,$$
$$\underline{m}_\Gamma(p,x) = \overline{w}\;\mathbb{E}[Y\,\mathbf{1}\{Y \leq \tilde{q}_\Gamma(x,p)\} \mid X = x, P = p] \;+\; \underline{w}\;\mathbb{E}[Y\,\mathbf{1}\{Y > \tilde{q}_\Gamma(x,p)\} \mid X = x, P = p]\,,$$

with $\tilde{q}_\Gamma(x,p)$ chosen analogously for the lower-tail allocation.[3]

Given these bounds on $m(p_L,x)$ and $m(p_H,x)$, the induced policy value bounds are simply

$$\underline{V}_\Gamma(\pi) = \mathbb{E}\left[\sum_{p\in\{p_L,p_H\}} \pi(p \mid X)\,(p-C)\,\underline{m}_\Gamma(p,X)\right], \qquad \overline{V}_\Gamma(\pi) = \mathbb{E}\left[\sum_{p\in\{p_L,p_H\}} \pi(p \mid X)\,(p-C)\,\overline{m}_\Gamma(p,X$$

These intervals are monotone in Γ and collapse to the standard plug-in estimand at $\Gamma = 1$. Economically, increasing Γ allows the latent state U to more strongly select into the observed price, which in turn permits more aggressive "adversarial" tilting of the outcome distribution within each price slice.

Alternative: additive bias bounds. Some organizations find it easier to reason about outcome-scale distortions than assignment odds ratios. A simple alternative postulates an additive deviation

$$|\mathbb{E}[Y(p) \mid X = x] - \mathbb{E}[Y \mid X = x, P = p]| \;\leq\; b(x,p),$$

leading immediately to $m(p,x) \in [\mathbb{E}[Y \mid X = x, P = p] - b(x,p),\, \mathbb{E}[Y \mid X = x, P = p] + b(x,p)]$ (clipped to $[0,\bar{y}]$), and hence linear bounds on $V(\pi)$. This specification is typically less sharp than odds-ratio models but can be more interpretable when business stakeholders can articulate "residual uplift" magnitudes.

Extension to multiple prices via linear programming. When $\mathcal{P}$ contains multiple price points, the same logic applies but the sharp identified set is most conveniently computed via optimization. For each $p \in \mathcal{P}$ and context cell (or learned representation) x, we treat $m(p,x)$ as an unknown decision variable. The sensitivity model induces linear constraints linking $(m(p,x))_{p\in\mathcal{P}}$ to observed conditional moments $\mathbb{E}[Y \mid X = x, P = p]$ through

[3] When Y has atoms at the threshold, one can randomize weights within the atom to satisfy $\mathbb{E}[w(Y)] = 1$ exactly; the resulting bounds remain sharp.

feasible reweightings (bounded by Γ and normalized within each (x, p) slice). The target policy value is linear in these means:

$$V(\pi) = \mathbb{E}\left[\sum_{p \in \mathcal{P}} \pi(p \mid X)\,(p - C)\,m(p, X)\right].$$

Thus, $\underline{V}_\Gamma(\pi)$ and $\overline{V}_\Gamma(\pi)$ are the solutions to a pair of linear programs (or convex programs, depending on how X is handled) obtained by minimizing/maximizing the above objective over the Γ-feasible set. Practically, this delivers a scalable procedure: we can report a sensitivity curve $\Gamma \mapsto [\underline{V}_\Gamma(\pi), \overline{V}_\Gamma(\pi)]$ and assess whether any decision-relevant comparisons between candidate pricing policies survive moderate levels of hidden confounding.

4.4 Minimal randomization as an instrument: a small exploration policy

Sensitivity bounds are useful when an organization is unwilling or unable to intervene in pricing, but they do not by themselves create new information: when the logged policy is highly endogenous, the identified set can remain wide even for moderate values of the sensitivity parameter. A complementary approach is to introduce *minimal* online randomization that is operationally acceptable yet sufficient to break the link between hidden state and price. Conceptually, we treat this randomization as an *instrument* that induces exogenous price variation while leaving the rest of the system unchanged.

Design: a Bernoulli exploration switch. We augment the production policy with a randomized switch $Z \in \{0, 1\}$. On each interaction (e.g., a page view, session, or pricing decision window), the system draws

$$Z \sim \mathrm{Bernoulli}(\rho),$$

possibly as a function of coarse observables (traffic tier, region) but, critically, independent of latent drivers of demand once we condition on X. When $Z = 0$ we serve the incumbent price generated by the existing logging policy, which may depend on unobserved factors:

$$P \mid (X, U, Z = 0) \;\sim\; \mu_0(\cdot \mid X, U).$$

When $Z = 1$ we *override* the incumbent decision and sample a price from a known exploration distribution:

$$P \mid (X, U, Z = 1) \;\sim\; g(\cdot \mid X),$$

where g is chosen by the practitioner and logged by construction. The exploration rate ρ can be small—often well below a few percent—so that the system remains primarily governed by the baseline policy, yet the exploration slice provides a source of randomized variation.

Two design principles matter. First, g must have support on the prices that the target policy π might choose:

$$\mathrm{supp}(\pi(\cdot \mid x)) \subseteq \mathrm{supp}(g(\cdot \mid x)) \quad \text{for all relevant } x.$$

Second, the assignment Z must be implemented in a way that is insulated from endogenous operational triggers (e.g., do not set $Z = 1$ only when inventory is high if inventory shocks are partly latent in U). In practice, we recommend generating Z from stable identifiers (user hash, request hash) and conditioning only on clearly observed strata included in X.

Identification: the exploration slice as a contextual experiment. The key causal insight is that, on $Z = 1$, price is randomized conditional on X, so it is independent of the unobserved confounder:

$$P \perp U \mid (X, Z = 1).$$

Under an exclusion/consistency condition stating that Z affects demand only through the posted price (no "experiment flag" effects, no change in ranking, shipping promises, or marketing exposure triggered by exploration), we can interpret the conditional mean outcome on the exploration slice as a causal estimand:

$$m(p, x) \;:=\; \mathbb{E}[Y(p) \mid X = x] \;=\; \mathbb{E}[Y \mid X = x, P = p, Z = 1],$$

for all (p, x) with $g(p \mid x) > 0$. This delivers point identification of the value of any target policy whose price support is covered by g:

$$V(\pi) \;=\; \mathbb{E}\left[\sum_{p \in \mathcal{P}} \pi(p \mid X)\,(p - C)\,m(p, X)\right],$$

with the integral form replacing the sum when prices are continuous. Economically, the exploration traffic creates a small, repeated randomized experiment embedded in production, converting an otherwise observational pricing system into one with a credible source of exogenous variation.

Local perturbations and guardrails. Operationally, organizations rarely permit arbitrary price randomization. More common is *local* exploration around a baseline price $p_0(X)$ (often the incumbent recommendation), subject to guardrails such as minimum margin, price floors/ceilings, and channel-specific constraints. A convenient specification is a discrete perturbation set

$$\mathcal{P}_{\mathrm{loc}}(x) \;=\; \{p_0(x) + \Delta_1, \ldots, p_0(x) + \Delta_K\} \cap \mathcal{P},$$

with $g(\cdot \mid x)$ supported on $\mathcal{P}_{\text{loc}}(x)$. This design can be made "safe" by choosing small perturbations and enforcing hard constraints deterministically (e.g., if a sampled price violates a compliance rule, resample within the feasible set). The tradeoff is conceptual rather than technical: local exploration identifies $m(p, x)$ only for those prices that are actually explored. Consequently, we can point-identify $V(\pi)$ only for target policies π that do not leave this explored neighborhood; evaluating more aggressive policy shifts then requires either (i) expanding the perturbation set over time, or (ii) invoking a structural or smoothness model to extrapolate beyond explored prices, which reintroduces modeling risk.

A further practical complication is *noncompliance*: downstream systems may override the randomized price due to last-minute constraints (stockouts, legal rules, competitor-matching). When this happens, Z is better interpreted as an *encouragement* rather than an assignment. The reduced-form randomization remains useful, but identification may shift from "intention-to-treat" effects to instrumental-variables estimands. In many pricing settings, however, it is feasible to engineer exploration so that compliance is near-perfect within well-chosen strata (e.g., only explore when inventory is ample and the price ladder is unconstrained), restoring the simpler identification above.

Operational feasibility and failure modes. Minimal randomization is often feasible because it can be implemented as a thin wrapper around the existing policy: sample Z, optionally sample an exploratory price, log (Z, g), and monitor. The principal cost is not computational but organizational: stakeholders must accept a controlled degree of short-run revenue risk in exchange for long-run learning. This is precisely why the parameter ρ is valuable: it gives a direct lever on how much traffic is exposed to experimental prices.

The main threats to validity are engineering and system interactions. First, the exclusion restriction can fail if setting $Z = 1$ changes more than price (e.g., the UI displays a "special offer" badge, the ranking algorithm responds to the price change in ways not attributable to the price itself, or marketing systems treat experimental sessions differently). Second, interference across units can arise if competitors react to exploratory prices, or if customers observe multiple prices over time and strategically delay purchases; such dynamics can be mitigated by randomizing at appropriate temporal or user-level clusters and by defining the outcome window to limit carryover. Third, randomization must be logged faithfully: without reliable records of Z and $g(p \mid X)$, the exploration slice loses its principal advantage—known propensities.

In sum, a small randomized exploration policy serves as an instrument that restores identifiability of counterfactual demand at explored prices, while remaining compatible with real-world guardrails. The remaining question is statistical and operational: given an exploration budget (or a revenue-loss cap), how should we estimate $V(\pi)$ efficiently and attach confidence intervals that reflect the fact that only a ρ fraction of the data are truly randomized. This is the focus of the next section.

4.5 Estimation and inference under known exploration propensities

Once we have engineered a slice of traffic in which the pricing propensities are known by design, the remaining difficulty is no longer causal identification but statistical efficiency and credible uncertainty quantification. The central object is the conditional mean demand surface identified on the exploration slice,

$$m(p, x) \ := \ \mathbb{E}[Y \mid X = x, P = p, Z = 1],$$

which delivers the value functional

$$V(\pi) \ = \ \mathbb{E}\left[\sum_{p\in\mathcal{P}} \pi(p \mid X)\,(p - C)\,m(p, X)\right].$$

In this section we describe estimators that (i) exploit the fact that $g(p \mid x)$ is known, (ii) remain valid under flexible outcome models, and (iii) admit finite-sample confidence intervals whose width makes the dependence on ρ operational.

Baseline estimators: inverse weighting and plug-in regression. A first approach is to use only the randomized observations and reweight them to match the target policy. Let $R_i := (P_i - C_i)Y_i$ denote profit and let $n_1 := \sum_{i=1}^{n} \mathbf{1}\{Z_i = 1\}$ be the number of exploration observations. For discrete $\mathcal{P}$, the exploration propensity satisfies $\mathbb{P}(P = p \mid X, Z = 1) = g(p \mid X)$, so an unbiased inverse-propensity estimator is

$$\widehat{V}_{\mathrm{IPW}}(\pi) \ := \ \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Z_i = 1\}\,\frac{\pi(P_i \mid X_i)}{g(P_i \mid X_i)}\,R_i,$$

with the obvious integral/density-ratio form when P is continuous. Intuitively, we “keep” only randomized interactions and upweight those whose realized price is likely under π but rare under g.

A complementary approach is to estimate $m(p, x)$ on the exploration slice and then plug it into the g-formula:

$$\widehat{V}_{\mathrm{REG}}(\pi) \ := \ \frac{1}{n}\sum_{i=1}^{n}\sum_{p\in\mathcal{P}} \pi(p \mid X_i)\,(p - C_i)\,\widehat{m}(p, X_i),$$

where $\widehat{m}$ may be obtained from any supervised learner trained on $\{(X_i, P_i, Y_i) : Z_i = 1\}$. This estimator uses the full sample to average over the marginal distribution of (X, C), but it inherits model bias if $\widehat{m}$ is misspecified.

Doubly robust (augmented) estimation on the exploration slice. To combine the robustness of reweighting with the stability of regression, we use an augmented inverse-weighting estimator (AIPW), specialized to the case of *known* propensities on $Z = 1$:

$$\widehat{V}_{\mathrm{DR}}(\pi) \;:=\; \frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_{p\in\mathcal{P}} \pi(p \mid X_i)\,(p - C_i)\,\widehat{m}(p, X_i) \\ +\; \mathbf{1}\{Z_i = 1\}\,\frac{\pi(P_i \mid X_i)}{g(P_i \mid X_i)}\,(P_i - C_i)\,\Big(Y_i - \widehat{m}(P_i, X_i)\Big)\Bigg].$$

The second term is a mean-zero correction when $\widehat{m}$ is accurate, which reduces sensitivity to regression error and typically lowers variance relative to pure IPW. The key practical point is that, because g is known and under our design is bounded away from zero on the relevant support, the instability that plagues off-policy evaluation with unknown (and potentially confounded) propensities is substantially mitigated.

For implementation with high-capacity learners, we recommend cross-fitting: partition the exploration slice into folds, estimate $\widehat{m}^{(-k)}$ on all but fold k, and evaluate the score on fold k. This controls overfitting bias while preserving the simplicity of the estimating equation.

Noncompliance and IV-style estimators. If exploration assigns a draw $\widetilde{P} \sim g(\cdot \mid X)$ but downstream constraints sometimes replace it with a served price P, then Z (or $\widetilde{P}$) becomes an encouragement rather than a treatment assignment. In that case the previous estimators remain valid for the *intention-to-treat* value of the exploration mechanism (the policy that sets Z and draws $\widetilde{P}$), but not automatically for the counterfactual value of a target policy on the served price. A tractable alternative is to impose a local structural restriction, e.g. a linear-in-price demand model on the exploration margin,

$$\mathbb{E}[Y \mid X, P] \approx \alpha(X) + \beta(X)\,P,$$

and estimate $\beta(X)$ using Z as an instrument for P (or using the randomized draw $\widetilde{P}$ as an instrument). This yields IV-style estimates of marginal price effects and, in turn, an approximate value calculation for policies that move prices within the exploration range. We view this as a useful fallback when perfect compliance is infeasible, but we emphasize the tradeoff: IV restores a form of identification by adding functional form.

Finite-sample confidence intervals and the role of ρ. Because only a fraction ρ of interactions are randomized, the effective sample size for causal learning is $n_1 \approx \rho n$. Under bounded outcomes $Y \in [0, \bar{y}]$, bounded margins $|P - C| \leq M$, and overlap within exploration $\inf_{x,p} g(p \mid x) \geq g_{\min} > 0$, the DR score is bounded in magnitude on $Z = 1$ by a constant of order $M\bar{y}/g_{\min}$. Standard concentration tools (e.g. Bernstein or empirical Bernstein inequalities applied to the cross-fitted score) therefore deliver a finite-sample interval

$$\mathrm{CI}_{1-\delta}(\pi) \;=\; \left[\widehat{V}_{\mathrm{DR}}(\pi) \;\pm\; c\, M \bar{y} \sqrt{\frac{\log(1/\delta)}{\rho n \, g_{\min}}} \right],$$

up to second-order nuisance-estimation terms that vanish under mild rates. This expression makes the operational comparative statics transparent: tightening the interval by a factor of two requires roughly four times as much randomized traffic (or four times the horizon), and poor overlap within g enters linearly through $1/\sqrt{g_{\min}}$ via the weight bound.

Choosing ρ under a revenue-loss cap. Exploration is ultimately governed by a budget constraint: stakeholders tolerate only so much short-run revenue risk. A simple planning rule is to translate that constraint into an upper bound on ρ and then check whether the resulting ρn suffices for statistical precision.

Let $p_0(X)$ be the incumbent (non-exploratory) price, and define the per-interaction profit loss from serving p instead of $p_0(X)$ as

$$\ell(p, X) \;:=\; (p_0(X) - C)\, m(p_0(X), X) \;-\; (p - C)\, m(p, X).$$

If we impose a conservative bound $\ell(p, X) \leq \ell_{\max}$ for all p in the exploration support (obtained from historical margins and a cautious elasticity envelope), then expected revenue loss per interaction from exploration is at most $\rho\, \ell_{\max}$. Over n interactions, a loss cap B implies

$$\rho \;\leq\; \frac{B}{n\, \ell_{\max}}.$$

Separately, to achieve a target half-width h at confidence level $1 - \delta$, the concentration bound suggests

$$\rho \;\geq\; \frac{c^2 M^2 \bar{y}^2 \log(1/\delta)}{n\, g_{\min}\, h^2}.$$

Feasibility therefore requires the interval $[\rho_{\min}, \rho_{\max}]$ to be nonempty; if it is empty, the remedy is mechanical: extend the horizon n, improve overlap by redesigning g (increase $g_{\min}$), relax h, or negotiate a larger exploration budget. In practice we often begin with a small ρ that comfortably satisfies the loss cap, use early data to tighten the bound on $\ell_{\max}$ (and to diagnose overlap), and then adjust ρ upward only if precision demands it. This makes the exploration rate a transparent policy lever: it directly prices the tradeoff between short-run revenue protection and long-run learnability.

4.6 Semi-synthetic experiments: stress-testing confounding, overlap, and exploration

Before committing to an online exploration rate ρ (and before trusting any particular estimator), we have found it useful to run *semi-synthetic* experiments. The goal is not to perfectly emulate consumer behavior, but to create a controlled environment in which (i) the empirical distribution of contexts X and costs C is realistic, (ii) confounding through an unobserved U is present by construction, and (iii) the "truth" $V(\pi)$ is known for a range of target policies. This lets us quantify, in the same units as deployment decisions (profit), how much bias arises from naive off-policy evaluation (OPE), how informative bounded-confounding intervals are, and how quickly minimal randomization corrects the problem as ρ increases.

Data backbone and injected confounders. We begin from a real log of contexts and operational signals, $\{(X_i, C_i)\}_{i=1}^{n}$, where X may include product attributes, calendar features, channel, and inventory indicators that are observable in production. We then inject an unobserved confounder U meant to represent one of three recurring sources of endogeneity in pricing systems: (i) *marketing shocks* (bursts in spend or impressions), (ii) *ranking/visibility changes* (algorithmic boosts that simultaneously increase demand and affect the pricing controller), and (iii) *stockout throttles* (latent inventory pressure that pushes price upward while mechanically suppressing sales). Operationally, we set $U_i \in \{0, 1\}$ (or a small discrete set) with $\mathbb{P}(U = 1 \mid X = x)$ chosen to match plausible seasonality and segmentation patterns; we then treat U as hidden at evaluation time.

A structural demand generator with tunable confounding. Given (X, U) and a posted price P, we generate demand from a bounded model so that profit remains well-scaled. For example, for purchase indicators one may use

$$\mathbb{P}(Y = 1 \mid X = x, P = p, U = u) \;=\; \sigma(s(x) \;-\; \eta\, p \;+\; \kappa\, u)\,,$$

where σ is the logistic link, $\eta > 0$ controls price sensitivity, and κ controls the strength of confounding (marketing/visibility shifts). For unit-demand counts one can instead generate $Y \in \{0, 1, \ldots, \bar{y}\}$ by thinning a Poisson mean and truncating at $\bar{y}$. We emphasize that the exact parametric form is not the point; what matters is that κ and the distribution of U give us an interpretable dial for how badly P is correlated with latent demand.

A confounded logging policy with realistic support. To mimic production logs, we generate prices from a logging policy that depends on U:

$$P \;\sim\; \mu_0(\cdot \mid X, U),$$

where μ_0 may be a discretized ladder with context-dependent support, reflecting guardrails and price floors/ceilings. A simple but revealing design is a mixture of two controllers: when $U = 1$ (high latent demand or marketing intensity), the system tends to post higher prices, while when $U = 0$ it posts lower prices; by varying the separation between these mixtures we can interpolate between mild endogeneity and the near-deterministic, no-overlap situation highlighted in Proposition 1. Limited support is imposed by context-specific allowed sets $\mathcal{P}(x) \subseteq \mathcal{P}$, so that some prices have $\mathbb{P}(P = p \mid X = x) = 0$ even when a target policy would assign them positive mass.

Target policies and "policy shift magnitude." We then define a family of target policies that gradually deviate from the logger. A convenient construction is

$$\pi_\lambda(\cdot \mid x) \;=\; (1 - \lambda)\, \widetilde{\mu}_0(\cdot \mid x) \;+\; \lambda\, \pi^\star(\cdot \mid x), \qquad \lambda \in [0, 1],$$

where $\widetilde{\mu}_0(\cdot \mid x)$ is the *observed* logging conditional (i.e., μ_0 marginalized over U) and $\pi^\star$ is an aspirational policy (e.g., a margin-maximizing rule or an RL policy trained offline). The parameter λ makes "how far we extrapolate from the logs" explicit; in practice it is a proxy for both overlap stress (mass on rarely seen prices) and the variance of reweighting-based estimators.

Estimators compared. On each semi-synthetic dataset we compute: (i) *Naive OPE* that treats the logs as unconfounded, e.g. IPW/DR using an estimated propensity $\widehat{\mu}(p \mid x)$ fit from (X, P) alone, or a direct regression plug-in using $\widehat{m}(p, x)$ fit from all data ignoring U. (ii) *Bounded-confounding intervals* $[\underline{V}_\Gamma(\pi), \overline{V}_\Gamma(\pi)]$ from Proposition 3, calibrated over a grid of Γ (and, when possible, anchored by domain knowledge about how strongly marketing or ranking can tilt price selection). (iii) *Minimal-randomization estimation* by synthetically adding an exploration slice: we draw $Z \sim \text{Bernoulli}(\rho)$ and, when $Z = 1$, resample $P \sim g(\cdot \mid X)$ with known propensities and full support over the prices needed by π. We then apply the estimator $\widehat{V}_{\text{DR}}(\pi)$ defined previously, using only the $Z = 1$ slice for causal identification (but averaging over all contexts).

What we typically observe. Across a wide range of calibrations, three empirical regularities recur. First, naive OPE can be *precisely wrong*: as κ increases, bias grows roughly linearly while estimated standard errors remain misleadingly small, because the estimated propensities $\widehat{\mu}(p \mid x)$ cannot account for the latent sorting on U. This is especially stark when the logging rule is nearly deterministic in U, where the estimator may confidently extrapolate into counterfactual strata with no information.

Second, Γ-bounds behave as intended: for small Γ they can be informative (and sometimes nearly point-like), while for large Γ they expand to reflect genuine ambiguity. In our experience, plotting interval width against Γ and λ is a useful diagnostic: rapid blow-up as λ increases is a concrete signal that a proposed policy relies on unsupported counterfactuals even under moderate confounding.

Third, the minimal-randomization estimator corrects bias quickly as soon as the exploration slice attains nontrivial effective coverage. When $g_{\min}$ is not too small, mean-squared error scales approximately like $1/(\rho n)$, matching the intuition from the finite-sample discussion. Conversely, when $g_{\min}$ is tiny (e.g. exploration puts very little mass on prices the policy cares about), variance dominates and the estimator becomes unstable even though it is, in principle, unbiased; this mirrors the practical importance of designing g for overlap rather than for "minimal perturbation" alone.

Sensitivity to overlap and policy shift. The interaction between overlap and policy shift is the main lesson for practice. Holding ρ fixed, increasing λ typically worsens performance in two ways: it increases reliance on high-variance weights $\pi(P \mid X)/g(P \mid X)$ and, if π assigns mass outside the exploration support, it creates an immediate identification failure. Semi-synthetic plots of error versus λ therefore double as a *policy feasibility check*: they reveal whether the policy class under consideration can be evaluated (and ultimately learned) under the intended exploration design.

Why we view these experiments as an engineering primitive. We do not claim that a semi-synthetic generator validates a pricing model; rather, it validates the *evaluation pipeline* under controlled violations that resemble production failure modes. The deliverable is a small set of empirically grounded design rules—minimal acceptable ρ, required $g_{\min}$, and a safe bound on policy shift λ—that can be communicated to stakeholders as a concrete instrumentation plan. These lessons feed directly into our next discussion of RL pricing: what cannot be learned from logs alone, and what minimal standards (randomization, logging, and guardrails) make deployment scientifically and economically safe.

Implications for RL pricing: what logs can and cannot support. The semi-synthetic results carry a direct message for RL-style pricing: without either (i) credible overlap or (ii) an instrument that creates an unconfounded slice, *the data do not determine the counterfactual.* This is not merely an estimator-choice problem. When the logging controller μ_0 adapts to latent demand shocks (our U), the apparent "state-action value" learned from logs conflates price effects with selection. In such settings, offline RL objectives (e.g. maximizing a learned $Q(x, p)$) optimize a quantity that is not causally meaningful; the resulting policy can look strong under retrospective evaluation yet underperform online. Conversely, when overlap fails—prices that the learned policy would choose are missing in parts of the context space—no amount of function approximation resolves the missing counterfactual support (Proposition 2). We therefore view the primary deliverable of an offline RL exercise as *a set of candidate policies whose value is either point-identified on a randomized slice or partially identified under explicit sensitivity parameters*, rather than a single "best" policy.

Sequential decision-making does not remove the identification bottleneck. In dynamic pricing, a natural hope is that long horizons and state augmentation (e.g. including inventory, lagged sales, or competitor features) render the problem observationally causal. Our framework suggests the opposite: adding state can help, but only insofar as it makes the relevant confounders observed. If unobserved forces remain (marketing bursts, ranking boosts, throttling logic, latent stock pressure), then the Markov decision process that the RL algorithm assumes is, from the evaluator's perspective, a partially observed process. The consequence is that standard off-policy evaluation identities for contextual bandits or MDPs (importance sampling, doubly robust temporal-difference estimators) inherit the same fragility: they require either ignorability (no residual U after conditioning) or an experimental lever that breaks the P–U link. Thus, "RL makes it dynamic" does not by itself produce causal identification; it simply spreads the confounding over time.

Recommended instrumentation standards for safe deployment. If we want RL claims (profit lift, fairness compliance) to be scientifically credible, we need to treat exploration and logging as first-class product requirements. In practice, we recommend an instrumentation bundle with three elements. First, *explicit randomization flags*: record Z indicating whether the served price came from controlled exploration, along with the realized propensities (the deployed $g(p \mid X)$, including any guardrail truncation). This is essential because, absent known propensities, one cannot reliably reconstruct weights, and the purported "randomized" slice becomes ambiguous. Second, *support design*: ensure $\operatorname{supp}(\pi(\cdot \mid x)) \subseteq \operatorname{supp}(g(\cdot \mid x))$ at the granularity at which decisions are made (product×channel×time-of-day, etc.). In pricing systems, silent guardrails (floors/ceilings, out-of-stock overrides, MAP constraints) often collapse support in precisely the contexts where policies differ; these must be logged as part of X or treated as regime changes. Third, *outcome alignment*: define Y and the decision window so that exclusion is plausible, i.e. Z affects Y only through P in that window. If exploration changes page layout, traffic mix, or latency, exclusion fails and the experimental slice no longer identifies $Y(p)$.

How much exploration is "enough" for RL? The correct benchmark is not a philosophical notion of exploration, but the sampling rate needed to bound decision-relevant error. With bounded outcomes and minimum propensity $g_{\min}$, Proposition 5 implies that for a fixed target policy the effective sample size is ρn and the uncertainty scales like $\widetilde{O}\big((\rho n g_{\min})^{-1/2}\big)$. For RL, one typically evaluates many candidate policies (or implicitly searches a policy class), so the operational requirement is stronger: we must either (i) allocate sufficient ρ to control a uniform deviation over the class, or (ii) restrict policy updates so that each new policy remains close to the support and weighting regime induced by g. This motivates *conservative* RL updates in pricing: impose a trust region constraint such as

$$\mathbb{E}[\mathrm{KL}(\pi_{t+1}(\cdot \mid X) \,\|\, g(\cdot \mid X))] \leq \tau,$$

or an explicit cap on importance ratios $\pi(P \mid X)/g(P \mid X)$, not as an optimization convenience but as an identification-and-variance constraint.

Safe policy improvement and robustness reporting. Even with instrumentation, we advocate reporting RL results in a form that reflects what is and is not learned. When identification relies on the randomized slice, we can provide point estimates and finite-sample confidence intervals for $V(\pi)$ and for group metrics such as $\Delta_{\text{price}}(\pi)$ or $\Delta_{\text{profit}}(\pi)$ using the same exploration-based estimators (compute group-conditional values on $Z = 1$). When only partial identification is available, we recommend publishing sensitivity curves $\Gamma \mapsto [\underline{V}_\Gamma(\pi), \overline{V}_\Gamma(\pi)]$ and making policy decisions via robust criteria (e.g. maximize the lower bound, or require $\underline{V}_\Gamma(\pi) - \overline{V}_\Gamma(\pi_0) \geq 0$ relative to a baseline π_0). This is the pricing analog of safe policy improvement: we do not deploy because an RL policy looks better on average, but because it is provably non-worse under transparent confounding budgets.

Extensions to multi-SKU pricing. Moving from a single SKU to a basket of K products replaces P with a vector $P \in \mathbb{R}^K$ and demand with a vector $Y \in \mathbb{R}^K$, with cross-elasticities and substitution. Two difficulties emerge. First is combinatorial support: even if each SKU has a modest ladder, $\mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_K$ is enormous, so overlap in the joint space is implausible. Second is multi-dimensional confounding: the same U (visibility, marketing) can shift the entire price vector and the entire demand system. The practical response is to instrument *structured* exploration. Rather than randomizing full price vectors, we randomize low-dimensional perturbations with known propensities (e.g. one SKU at a time, or a small set of tagged

SKUs) while holding others at baseline. Identification then targets marginal effects that can be composed under modeling assumptions (e.g. sparsity or low-rank cross-effects). Formally, we can let Z index which coordinate (or group) is perturbed, and require $P_j \perp U \mid (X, Z = j)$ for the randomized coordinate; this yields credible causal estimates for partial derivatives or local treatment effects that can be used inside a constrained optimizer.

Competitive settings and interference. In many categories, price changes affect competitors, and competitor actions feed back into demand. This creates interference: one unit's outcome depends on other units' treatments, violating the exclusion/consistency conditions needed for Proposition 4 if exploration changes the market environment. A minimal way to restore credibility is to randomize at a level where interference is approximately contained (e.g. geo-level experiments, time-block randomization, or customer-level holdouts with minimal spillovers), and to redefine X to include the competitive state observed at decision time. Even then, policy evaluation becomes equilibrium-sensitive: $V(\pi)$ depends on how rivals respond. For deployment, we therefore recommend distinguishing *short-run causal effects holding competitors fixed* (identified by brief, small-ρ perturbations) from *long-run strategic effects* (requiring either longer experiments or structural modeling). RL systems that ignore this distinction may learn policies that exploit transient competitor inertia but fail once competitors adapt.

Bottom line for practice. RL pricing is feasible and valuable, but only when paired with instrumentation that makes causal evaluation routine: explicit randomization, logged propensities, engineered overlap, and monitoring of guardrail-induced support shifts. Where those conditions are not met, we can still proceed—but the correct object is a robustness analysis and a conservative deployment rule, not a single-number claim of profit lift.

5 Conclusion and policy/audit checklist: minimal requirements for credible claims

Our central conclusion is practical: pricing logs are not, by default, evidence about counterfactual profit or fairness. When historical prices are chosen in response to latent forces that also move demand, or when candidate policies place mass on rarely (or never) observed prices in parts of the context space, offline evaluation becomes an exercise in extrapolation rather than measurement. The economic lesson is that the bottleneck is not computational sophistication but the informational content of the data-generating process. Credible claims therefore require an explicit "causal contract" between the product system and the evaluator: what was randomized, with what probability, and under what outcome definition.

We summarize this contract as a minimum viable checklist for two classes of claims: (i) *profit lift* relative to a baseline policy, and (ii) *fairness* (price or profit differences across protected or business-critical groups). The checklist is intentionally auditable: each item can be verified from logged artifacts and simple diagnostics rather than from model fit alone.

A. Minimum requirements for a profit-lift claim. A profit-lift claim is credible only if the evaluation data identify the relevant causal demand objects at the prices the target policy would use. In deployment terms, this requires:

- **Explicit assignment metadata.** Every record must indicate whether the served price came from controlled randomization, and must include the realized propensity for that price under the randomization protocol (including any truncation due to floors/ceilings, inventory overrides, or compliance rules).
- **Designed support for the intended use.** For each context granularity at which decisions are effectively made (e.g. SKU×channel×time block), the randomized component must place positive probability on all prices (or price bins) that the target policy may select. If the system silently collapses support in certain contexts, the evaluation must either (i) restrict the policy to the supported region, or (ii) treat those contexts as outside scope.
- **Outcome-window integrity.** The sales outcome and its time window must be defined so that the randomization affects demand only through the posted price in that window. If randomization changes other primitives (ranking, page composition, latency, traffic allocation), then the evaluation no longer corresponds to a price experiment and must be reframed accordingly.
- **Effective sample size accounting.** The evaluation report must state the randomized traffic share and the minimum propensity (or an empirical lower tail) to make clear the variance implications of weighting. A point estimate without an uncertainty statement is not a lift claim; it is a hypothesis.
- **Stability and drift checks.** Because pricing systems operate in nonstationary environments, the report must include a time-split analysis (or rolling window) demonstrating that the estimated effect is not driven by a transient regime (promotion weeks, supply disruptions, competitor shocks). When drift is material, the claim should be localized in time and scope.

A useful internal standard is that the final lift statement be phrased as an interval (or a lower confidence bound) tied to a clearly specified policy and population, rather than as a single "expected lift" number. This forces discipline about what is identified and what is extrapolated.

B. Minimum requirements for a fairness claim. Fairness claims are stricter than profit claims because they are inherently subgroup claims: they require identification *within* groups and are sensitive to sparse support. We recommend the following additional requirements:

- **Group definitional clarity.** The group attribute(s) used for auditing must be defined, versioned, and linked to the decision-time features. If group labels are inferred or updated asynchronously, the analysis must document the labeling lag and error.
- **Within-group overlap and propensities.** For each group (and, ideally, for major intersections), the randomized component must cover the prices the evaluated policy would assign in that group. Otherwise, a fairness estimate is dominated by unsupported counterfactuals.
- **Group-conditional uncertainty.** Report confidence intervals (or bounds) for group-conditional values and for the chosen disparity metric (price disparity, profit disparity, or welfare proxies). If many groups are examined, control familywise error or report multiplicity-adjusted uncertainty.
- **Guardrail heterogeneity audit.** Verify that constraints (minimum advertised price, stockout throttles, eligibility rules) do not differentially bind across groups in a way that mechanically induces price differences. When they do, the fairness question should be reframed as a constraint-design question, not an algorithm-performance question.

In our experience, the most common fairness failure mode is not that a model "chooses to discriminate," but that operational constraints and missing support make subgroup effects unlearnable while still allowing optimistic aggregate summaries.

C. A practical audit workflow. To make these requirements operational, we suggest a three-stage evaluation workflow that separates *data adequacy* from *estimation*:

1. **Design audit (before learning).** Enumerate the target policy class and verify that the randomization protocol covers its action support in the relevant context strata; set explicit propensity floors.

2. **Logging audit (during deployment).** Monitor realized propensities, the randomized traffic share, and the frequency of guardrail overrides; alert when effective overlap deteriorates.

3. **Causal audit (after data collection).** Estimate policy value (and group metrics) using methods that respect the randomization design; report confidence intervals or, when relying on sensitivity models, report robustness curves indexed by a clearly interpretable confounding budget.

This workflow has a governance advantage: it yields pass/fail criteria that product, engineering, and risk teams can jointly own, rather than delegating credibility to a modeling team ex post.

Limitations. Our framework does not eliminate all ambiguity. First, minimal randomization identifies effects for the explored price distribution and time window; large policy shifts still require either substantial exploration or additional structure. Second, interference and equilibrium feedback remain fundamental obstacles in competitive markets; a clean price experiment at the user level need not identify long-run outcomes once competitors respond. Third, multi-SKU settings introduce high-dimensional action spaces where full-support exploration is infeasible; practical identification will often be local and modular rather than global. Finally, the approach presumes that the instrumentation itself is faithfully implemented; mis-logged propensities or unrecorded overrides can quietly reintroduce confounding.

Future work. Several directions are immediate. On the design side, we need exploration schemes that target identification efficiency under business constraints (inventory risk, margin floors), including adaptive designs that preserve known propensities. On the inference side, extending finite-sample guarantees to settings with interference, continuous prices, and heavy-tailed outcomes remains important. On the governance side, we view fairness under partial identification as underdeveloped: robust decision rules that trade off profit and disparity under transparent uncertainty budgets are a natural next step. More broadly, we expect credible RL pricing to converge toward a hybrid discipline: optimization guided by models, but claims anchored in explicitly randomized (or explicitly bounded) causal evidence.