Governed GP-UCB for Dynamic Pricing: Regret Guarantees under Fairness and Volatility Constraints

Liz Lemma Future Detective

January 16, 2026

Abstract

Dynamic pricing systems in 2026 face an additional objective beyond revenue: governance. Firms increasingly must respect fairness rules (e.g., non-discrimination across protected groups) and operational or regulatory limits on price volatility, while demand remains unknown and complex. Building on BO/GP-bandit pricing methods that avoid parametric demand assumptions, we propose Governed GP-UCB: a constrained Bayesian optimization algorithm that learns both revenue and constraint functions using Gaussian process posteriors. The method selects prices by maximizing an optimistic revenue bound subject to conservative feasibility bounds, ensuring high-probability constraint satisfaction (safe set expansion) or controlled violations (primal–dual). To address real deployment requirements, we incorporate explicit price-volatility limits via batching (hard switch budgets) or switching-cost regularization. We provide regret bounds comparable to unconstrained GP-UCB up to constraint-dependent terms and show how governance tightness shapes optimal prices through KKT characterizations in a clean static benchmark. Empirically, the governed algorithm achieves strong revenue while producing fewer price changes and substantially reduced disparate-impact measures, compared to unconstrained BO and deep RL baselines. The paper provides a practical, theory-backed blueprint for compliance-by-design pricing systems.

Table of Contents

1. 1. Introduction and motivation: governance constraints in modern pricing (fairness, volatility), gaps in parametric/RL methods, and why GP/BO is a natural compliance-by-design tool.

2. 2. Related work: dynamic pricing with learning (parametric/RL), GP bandits/BO, constrained/safe bandits, fairness in online decision-making, bandits with switching costs.

3. 3. Model: groups, demand/revenue primitives, observable feedback, constraint definitions (price parity and outcome-based fairness), and volatility constraints (switch/TV). Define feasible comparator classes.

4. 4. Static benchmark and characterization: constrained revenue maximization problem, KKT conditions, and interpretable comparative statics (how tighter fairness/volatility changes the optimal pricing vector). Flag when numerical optimization is needed.

5. 5. Algorithm: Governed GP-UCB

6. 5.1. GP surrogates for revenue and constraints; confidence sets and information gain.

7. 5.2. Safe action set construction (conservative constraints) and safe set expansion.

8. 5.3. Primal–dual extension when safety constraints are soft (bounded violation).

9. 5.4. Volatility control: (i) batching for hard switch budgets; (ii) regularized selection for switching costs/TV constraints.

10. 6. Theory

11. 6.1. High-probability feasibility (always-safe constraints) or sublinear cumulative constraint violation (primal–dual).

12. 6.2. Regret bounds vs best feasible policy in the volatility class; discuss dependence on number of groups and constraints.

13. 6.3. Discussion of kernel misspecification and practical hyperparameter tuning.

14. 7. Experiments and evaluation protocol

15. 7.1. Synthetic environments (nonstationary, non-monotone) with fairness ground truth.

16. 7.2. Semi-synthetic marketplace data: offline replay + safe online simulation.

17. 7.3. Metrics: revenue, regret proxies, fairness violations, number of price changes, stability, and auditability.

18. 8. Policy and deployment implications: audit logs from GP posteriors, interpretability, and how to choose fairness/volatility parameters in practice.

19. 9. Conclusion and extensions: contextual pricing, competition, and discrete demand likelihoods.

1 1. Introduction and motivation: governance constraints in modern pricing (fairness, volatility), gaps in parametric/RL methods, and why GP/BO is a natural compliance-by-design tool.

Modern pricing is increasingly governed as much by *constraints* as by demand. Platforms that can tailor offers at the level of protected groups (and, in practice, even individuals) face external scrutiny from regulators, internal risk teams, and consumer advocates. Two classes of restrictions recur across domains. First are *fairness* and non-discrimination requirements, which can take the form of direct limits on price dispersion (e.g., $|p(g) - p(g')| \leq \Delta$) or limits on disparities in predicted outcomes (e.g., $|\mu_g(p(g)) - \mu_{g'}(p(g'))| \leq \varepsilon$). Second are *operational volatility* constraints: frequent price changes can trigger consumer backlash, complicate communication, and violate contractual or policy commitments, motivating explicit budgets on the number of switches or on total variation over time. These governance requirements turn dynamic pricing into a sequential decision problem in which the seller must learn demand while remaining compliant period by period.

The methodological gap is that much of the dynamic-pricing literature optimizes within either parametric demand families or reinforcement-learning architectures that prioritize asymptotic performance under stationarity and rich exploration. Parametric models deliver interpretability and statistical efficiency when correctly specified, but they are brittle when group-level demand exhibits nonlinearities, thresholds, or unmodeled interactions. In governed settings, misspecification is not merely a welfare loss; it can induce systematic constraint violations (e.g., underestimating the demand response in one group leads to persistent outcome disparities) or over-correction that sacrifices revenue unnecessarily. At the other extreme, model-free RL is often paired with aggressive experimentation that is hard to reconcile with compliance: without calibrated uncertainty, exploration can wander into unsafe regions, and the resulting policies can be difficult to audit or justify ex post to regulators and stakeholders.

We therefore take a compliance-by-design perspective: the learning algorithm should explicitly track uncertainty about both *revenue* and *constraints*, and should select prices only when it can certify feasibility (or, under a softer regime, can quantify and control expected violations). Gaussian process (GP) models, used as nonparametric priors over unknown functions, naturally instantiate this idea. By producing posterior means and confidence bands, a GP approach converts unknown constraints into *probabilistic safety envelopes*. In particular, instead of hoping that a learned policy satisfies fairness constraints on average, we can require that an upper confidence bound on each constraint be nonpositive, ensuring that the chosen action is conservative relative to estimation error. The same uncertainty quantification supports principled exploration via optimism: upper confidence bounds for the objective encourage learning where revenue is potentially high, but only within the region currently certified as feasible.

This GP-based viewpoint is also well matched to the economic structure of governed pricing. Governance constraints couple groups: a parity limit ties prices directly across g, while outcome-based constraints link prices through demand responses. Such coupling makes it difficult to decompose learning group by group, and it complicates ad hoc exploration strategies. A joint Bayesian surrogate for revenue and constraints provides a unified representation of these cross-group tradeoffs and their uncertainty. Moreover, volatility constraints fit naturally into this framework as restrictions on the admissible sequence of actions (e.g., piecewise-constant prices with limited switches, or a total-variation budget), which can be enforced mechanically alongside feasibility. In operational terms, this yields policies that are not only statistically grounded but also implementable: they change prices infrequently, document why a change is warranted, and maintain explicit buffers against constraint violations.

We emphasize limitations alongside the appeal. GP models rely on regularity assumptions (captured here by an RKHS norm bound) that may fail under sharp discontinuities or strategic consumer responses. Constraint feedback may be indirect, noisy, or delayed, especially for outcome-fairness notions that depend on latent conversion probabilities rather than directly observed demand. Finally, conservative safety can be costly early on when uncertainty is large, potentially shrinking the set of allowable prices. Nonetheless, for the central governance problem—learn while remaining compliant—the GP/BO toolkit provides an unusually coherent combination of (i) flexible nonparametric approximation, (ii) calibrated uncertainty, and (iii) algorithmic machinery for safe, constrained optimization under sequential data.

2 2. Related work: dynamic pricing with learning (parametric/RL), GP bandits/BO, constrained/safe bandits, fairness in online decision-making, bandits with switching costs.

A large literature studies dynamic pricing with learning under parametric demand models. Classic approaches posit a demand curve indexed by an unknown parameter (often linear or logit), and use bandit-style experimentation to learn the parameter while controlling revenue loss; see, e.g., **???**. These models yield sharp regret bounds and transparent comparative statics when correctly specified, and they align naturally with managerial forecasting pipelines. Their limitation for our purposes is that the governance object is typically not modeled as an explicit constraint: fairness or parity restrictions, when present at all, are imposed ex post through ad hoc adjustments. Moreover, when parametric misspecification is present, the resulting bias can interact with group heterogeneity in precisely the way regulators care about (systematic disparities rather than purely noisy errors).

At the other end, reinforcement-learning formulations treat pricing as a control problem with rich state dynamics (inventory, demand seasonality, competition). This line is well suited to nonstationary environments and to settings with delayed rewards; see surveys such as **?**. However, RL methods commonly rely on exploration heuristics whose safety properties are difficult to certify, and constraint handling is often heuristic (reward shaping, Lagrangian penalties without calibrated uncertainty). In governance contexts, where the relevant question is frequently "was each decision compliant at the time it was made?", this gap between asymptotic performance and period-by-period guarantees is consequential.

Our methodological building block comes from Gaussian-process bandits and Bayesian optimization, which provide nonparametric function learning together with explicit posterior uncertainty. The GP-UCB paradigm **?** and its many variants establish regret rates that scale with kernel complexity via information gain, and have become standard tools for optimization under expensive, noisy evaluations. While most BO work targets low-frequency experimentation (e.g., hyperparameter tuning), the conceptual match to pricing is the combination of flexible response surfaces and calibrated confidence sets.

A closely related stream studies constrained and "safe" bandits, in which actions must satisfy unknown constraints learned from data. Safe BO methods construct conservative feasible sets from upper confidence bounds on constraint functions, ensuring feasibility with high probability; see, e.g., **??**. Constrained contextual bandits and primal–dual online learning methods provide complementary perspectives in which feasibility is traded off against reward via dual variables, yielding sublinear cumulative violations under suitable conditions **??**. Our setting differs in two respects that matter for practice: (i) constraints are governance constraints that couple protected groups through parity or outcome-based requirements, and (ii) the action is a *vector* of group prices, so both objective and constraints live on a joint price space rather than decomposing cleanly by group.

We also connect to the rapidly growing literature on fairness in online decision-making. Much work focuses on allocation or classification under statistical notions of fairness (e.g., equal opportunity) and studies the exploration–fairness tradeoff when group labels are observed **??**. Pricing raises distinct issues: the fairness object may be the price itself (disparate treatment) or the induced purchase outcomes (disparate impact), and both notions are naturally expressed as constraints on a decision-dependent demand system. Our formulation treats these requirements symmetrically as unknown functions to be learned and controlled, which facilitates auditing:

the same confidence machinery that justifies exploration also explains why a price vector is deemed compliant.

Finally, we relate to bandits with switching costs and variation budgets, which formalize the operational friction from frequent policy changes ?. In pricing, such constraints capture menu costs, customer trust, and internal governance processes (approvals, documentation). Technically, switching constraints interact with learning because they limit how quickly the algorithm can correct mistaken inferences. Our approach accommodates these frictions by restricting the admissible *sequence* of price vectors (hard switch or total variation), bringing together safe exploration, fairness governance, and operational stability in a single sequential decision framework.

3 3. Model: groups, demand/revenue primitives, observable feedback, constraint definitions (price parity and outcome-based fairness), and volatility constraints (switch/TV). Define feasible comparator classes.

We model a seller (or platform) interacting with consumers partitioned into a finite set of protected groups $\mathcal{G}$ with $G = |\mathcal{G}|$. Time is discrete, $t = 1, \ldots, T$. In each period the seller posts a *vector* of group-dependent prices $p_t \in [p_\ell, p_h]^G$, where $p_t(g)$ denotes the price offered to group g. The key economic primitive is that groups may respond differently to price due to heterogeneous elasticities, baseline willingness-to-pay, or differential access, so our action is intrinsically multidimensional.

When group g is offered price p, demand is a random variable $D_{g,t}(p)$ (units purchased, conversions, or accepted offers). We write the mean demand curve as

$$\mu_g(p) \;=\; \mathbb{E}[D_{g,t}(p)]\,,$$

and define realized and mean revenues for group g as

$$r_{g,t} \;=\; p_t(g)\,D_{g,t}\big(p_t(g)\big), \qquad f_g(p) \;=\; p\,\mu_g(p).$$

Total mean revenue from a price vector p is therefore

$$F(p) \;=\; \sum_{g\in\mathcal{G}} f_g\big(p(g)\big),$$

and the seller's realized revenue in period t is $\sum_g r_{g,t}$. To isolate the governance–learning tradeoff, our clean baseline assumes that after posting p_t the seller observes group-level revenues $r_{g,t}$ (or equivalently demands $D_{g,t}$), with conditionally σ-sub-Gaussian noise around the mean. In many applications one observes only aggregated outcomes or delayed chargebacks; these features can be incorporated, but they obscure the core mechanism we wish to emphasize: governance turns pricing into constrained optimization on a joint price space.

Governance enters through explicit constraints on the posted price vector. We encode each requirement as an (unknown) constraint function $c_j : [p_\ell, p_h]^G \to \mathbb{R}$ with the feasibility condition $c_j(p) \le 0$, for $j = 1, \ldots, J$. This abstraction allows the constraint to be a direct legal rule (e.g., disparate treatment in prices) or an operationalized policy threshold derived from internal audits. Two canonical examples are:

$$\text{(price parity)} \qquad c^{par}_{g,g'}(p) \;=\; |p(g) - p(g')| - \Delta \;\le\; 0,$$

which limits dispersion in treatment across groups up to tolerance Δ, and

$$\text{(outcome fairness)} \qquad c^{out}_{g,g'}(p) \;=\; \big|\mu_g(p(g)) - \mu_{g'}(p(g'))\big| - \varepsilon \;\le\; 0,$$

which limits disparities in *expected* purchase outcomes up to tolerance ε. The latter is economically natural but informationally demanding: it couples groups through unknown demand curves, so compliance requires learning how each group responds to its offered price. In practice, platforms may estimate these constraints via noisy proxies (conversion rates, acceptance probabilities, or complaint-adjusted outcomes); our formulation treats such measurements as noisy feedback on $c_j(p_t)$.

Finally, we impose volatility constraints capturing menu costs, approval frictions, and the governance reality that frequent price changes are themselves risky or infeasible. We study two common forms. A hard switch budget S limits the number of periods in which the price vector changes,

$$\#\{t \ge 2 : p_t \ne p_{t-1}\} \;\le\; S,$$

while a total-variation budget V limits cumulative movement,

$$\sum_{t=2}^{T} \|p_t - p_{t-1}\|_1 \;\le\; V.$$

Both restrictions force the learning algorithm to "live with" earlier choices, which is particularly consequential when constraints are present: the cost of a misstep is not merely low revenue, but potential noncompliance.

These elements jointly define the policy classes against which we evaluate performance. Let Π_S denote the class of feasible price sequences that satisfy all governance constraints period-by-period and incur at most S switches:

$$\Pi_S \;=\; \Big\{\{p_t\}_{t=1}^T : p_t \in [p_\ell, p_h]^G,\; c_j(p_t) \le 0\,\forall j, t,\; \#\{t \ge 2 : p_t \ne p_{t-1}\} \le S\Big\}.$$

Analogously, Π_V imposes the total-variation budget V. These comparator classes formalize a practical benchmark: we do not compare a governed, operationally constrained seller to an unconstrained clairvoyant, but to the best *feasible* pricing path within the same governance and stability requirements.

4 4. Static benchmark and characterization: constrained revenue maximization problem, KKT conditions, and interpretable comparative statics (how tighter fairness/volatility changes the optimal pricing vector). Flag when numerical optimization is needed.

Before turning to learning dynamics, we anchor the analysis in a *static governed benchmark*: what a regulatorily compliant, fully informed seller would do in a single period absent estimation error. This benchmark isolates the purely economic distortion induced by governance, separating it from the additional distortion created by learning under uncertainty.

Formally, given mean revenue $F(p)$ and governance constraints $\{c_j(p) \leq 0\}_{j=1}^{J}$, the static problem is

$$p^\star \in \arg \max_{p \in [p_\ell, p_h]^G} F(p) \qquad \text{s.t.} \qquad c_j(p) \leq 0, \quad j = 1, \ldots, J.$$

This formulation is deliberately agnostic about the origin of c_j: some constraints are *directly* about treatment (e.g., parity of posted prices), while others are *outcome-based* and therefore couple prices through demand responses. Economically, the distinction matters because treatment constraints primarily restrict the *geometry* of feasible prices, whereas outcome constraints restrict prices through the slope and curvature of the demand system, effectively importing behavioral heterogeneity into compliance.

When F is concave and each c_j is convex (as in convex surrogates for absolute-value fairness rules), the benchmark admits a transparent first-order characterization. There exist multipliers $\lambda_j^\star \geq 0$ such that

$$\nabla F(p^\star) + \sum_{j=1}^{J} \lambda_j^\star \nabla c_j(p^\star) = 0, \qquad \lambda_j^\star c_j(p^\star) = 0, \qquad c_j(p^\star) \leq 0.$$

The economic content is that governance introduces shadow prices: $\lambda_j^\star$ measures the marginal revenue gained from relaxing constraint j by one unit. In particular, under price-parity rules (bounding $\max_g p(g) - \min_g p(g)$), the multipliers encode how costly it is to prevent the platform from tailoring markups to group-specific elasticities. Under outcome-fairness rules, the multipliers instead price the induced coupling across groups: compliance requires shifting prices away from each group's unconstrained revenue optimum until expected outcomes align within tolerance.

This benchmark yields interpretable comparative statics that connect directly to policy levers. If Δ is the allowed parity dispersion, increasing Δ weakly expands the feasible set, implying

$$F\big(p^\star(\Delta_2)\big) \;\geq\; F\big(p^\star(\Delta_1)\big) \qquad \text{for} \quad \Delta_2 \geq \Delta_1,$$

and (generically) permits greater cross-group price dispersion. Similarly, increasing the outcome-fairness tolerance ε weakly raises the optimal governed value, reducing the extent to which high-elasticity groups "pull" other groups' prices through coupled constraints. Operationally, these monotonicities clarify what is being traded off when compliance teams debate tolerances: tighter rules reduce legal or reputational exposure but impose an endogenous tax on price discrimination that is often largest precisely when groups differ most.

Volatility constraints can be interpreted as adding *intertemporal* feasibility that, period-by-period, resembles a local restriction around the status quo. For example, with a per-period TV move limit v_t, the seller effectively solves

$$\max_p F(p) \quad \text{s.t.} \quad c_j(p) \le 0 \ \forall j, \quad \|p - p_{t-1}\|_1 \le v_t,$$

so tightening v_t (or reducing the switch budget S in a piecewise-constant regime) shrinks the feasible correspondence and can force persistence at suboptimal but "approved" prices. In KKT terms, an additional multiplier on the movement constraint captures the marginal value of flexibility, which is exactly what operational teams experience as the cost of slow price-approval pipelines.

A limitation of the static characterization is that it is only as tractable as the underlying geometry. Absolute-value fairness constraints, unknown demand-induced outcomes, and high-dimensional p can easily break convexity or differentiability, in which case closed-form characterization is unavailable and numerical methods are required even with full information. In our setting the challenge is sharper: F and c_j are *unknown* and must be learned. This is why we next turn to a GP-based procedure that simultaneously searches for high revenue and enforces conservative feasibility while respecting volatility.

5 5. Algorithm: Governed GP-UCB

We model both revenue and governance objects with Gaussian-process (GP) surrogates, which provide a disciplined way to translate finite, noisy observations into (i) point predictions and (ii) uncertainty quantification that can be carried forward into conservative compliance. Concretely, for each group $g \in \mathcal{G}$ we treat the mean revenue curve $f_g : [p_\ell, p_h] \to \mathbb{R}$ as an unknown function in the RKHS $\mathcal{H}_k$ with $\|f_g\|_{\mathcal{H}_k} \le B$. We place the corresponding GP prior $f_g \sim \mathcal{GP}(0, k)$ (or interpret the GP posterior as a kernel ridge estimator with calibrated uncertainty), and we update this posterior using the realized group revenues

$$y^F_{g,t} \;=\; r_{g,t} \;=\; p_t(g)\, D_{g,t}\big(p_t(g)\big) \;=\; f_g\big(p_t(g)\big) + \epsilon_{g,t}, \qquad \epsilon_{g,t} \text{ is } \sigma\text{-sub-Gaussian.}$$

Given data $\mathcal{D}^F_t = \{(p_\tau(g), y^F_{g,\tau})\}_{\tau \le t,\, g \in \mathcal{G}}$, standard GP regression yields, for each group, a posterior mean $\mu^F_{g,t}(\cdot)$ and posterior standard deviation $\sigma^F_{g,t}(\cdot)$.

Because the seller's objective is the sum across groups, we propagate these group-level posteriors to total mean revenue

$$F(p) \;=\; \sum_{g\in\mathcal{G}} f_g\big(p(g)\big), \qquad \mu_t^F(p) \;=\; \sum_{g\in\mathcal{G}} \mu_{g,t}^F\big(p(g)\big),$$

and we upper bound uncertainty in F by aggregating group uncertainties (e.g., via a union bound across groups), which is what ultimately permits a single optimistic score for any candidate price vector $p \in [p_\ell, p_h]^G$.

We treat governance constraints analogously. For each constraint $j = 1, \dots, J$, we posit an unknown function $c_j : [p_\ell, p_h]^G \to \mathbb{R}$ with $\|c_j\|_{\mathcal{H}_k} \le B_c$ and noisy feedback

$$y_{j,t}^c \;=\; c_j(p_t) + \eta_{j,t}, \qquad \eta_{j,t} \text{ is } \sigma_c\text{-sub-Gaussian.}$$

Here $y_{j,t}^c$ can be a direct measurement (e.g., a computed parity gap) or an estimable proxy (e.g., an outcome disparity inferred from conversion estimates); the GP abstraction simply requires that we can form a scalar observation whose conditional mean is $c_j(p_t)$. Updating on $\mathcal{D}_t^c = \{(p_\tau, y_{j,\tau}^c)\}_{\tau\le t}$ yields $\mu_{j,t}^c(\cdot)$ and $\sigma_{j,t}^c(\cdot)$.

The key deliverable from these posteriors is a sequence of high-probability confidence bands. For revenue we form

$$U_t^F(p) \;=\; \mu_t^F(p) \;+\; \sqrt{\beta_{t+1}}\;\tilde{\sigma}_t^F(p), \qquad L_t^F(p) \;=\; \mu_t^F(p) \;-\; \sqrt{\beta_{t+1}}\;\tilde{\sigma}_t^F(p),$$

and for each constraint,

$$U_{j,t}^c(p) \;=\; \mu_{j,t}^c(p) \;+\; \sqrt{\beta_{t+1}}\;\sigma_{j,t}^c(p), \qquad L_{j,t}^c(p) \;=\; \mu_{j,t}^c(p) \;-\; \sqrt{\beta_{t+1}}\;\sigma_{j,t}^c(p).$$

The exploration scale β_t is chosen so that, with probability at least $1-\delta$, these bands hold uniformly over t and p (and over all g, j); operationally, this is the statistical content behind conservative compliance: we certify feasibility using an upper bound U^c, rather than a point estimate that may be wrong early on.

Finally, our regret and sample-complexity accounting is organized by the GP *information gain*,

$$\Gamma_T \;=\; \max_{A:\,|A|=T} \; \frac{1}{2}\log\det\Big(I + \sigma^{-2}K_A\Big),$$

where K_A is the kernel Gram matrix on the queried price points. Γ_T measures how quickly posterior uncertainty can shrink under kernel k; it is the bridge from per-period confidence widths to cumulative learning performance. Economically, it captures the effective complexity of the demand-and-governance environment: smoother functions (under k) generate smaller Γ_T, meaning fewer exploratory price experiments are needed to both learn and remain compliant. A limitation, which we treat as a modeling caveat rather than a technicality, is that misspecified kernels or nonstationary behavior can inflate realized uncertainty relative to the nominal GP bands, motivating diagnostics and conservative choices of k and β_t in practice.

6 5.1. GP surrogates for revenue and constraints; confidence sets and information gain.

5.2. Safe action set construction (conservative constraints) and safe set expansion. The confidence bands from Section 5.1 become operational only once we convert them into an *admissible* set of price vectors that we are willing to deploy online. Our guiding principle is simple: when constraints represent governance rules (parity, outcome fairness, regulatory caps), we prefer *ex ante* certification to *ex post* repair. In the GP language, this means we enforce constraints using an upper confidence bound, so that any action deemed feasible is feasible for the true (unknown) constraint with high probability.

Formally, at the start of period t we construct the conservative feasible region

$$\mathcal{A}_t^{safe} \;=\; \Big\{p \in [p_\ell, p_h]^G : \; U^c_{j,t-1}(p) \le 0 \;\; \forall j = 1, \dots, J\Big\},$$

where $U^c_{j,t-1}(p) = \mu^c_{j,t-1}(p) + \sqrt{\beta_t}\sigma^c_{j,t-1}(p)$ is the period-$(t-1)$ upper band for constraint j. The economic interpretation is that we price as if the constraint were at its *most pessimistic* level consistent with observed data: only if even the pessimistic assessment is compliant do we treat p as implementable. Under the same event on which the GP bands are valid uniformly over time and actions, any realized choice $p_t \in \mathcal{A}_t^{safe}$ satisfies $c_j(p_t) \le 0$ for all j, which is the sense in which conservative feasibility turns statistical uncertainty into a governance guarantee.

Two practical issues arise immediately. First, $\mathcal{A}_t^{safe}$ must be nonempty. In applications this is typically handled by positing (or eliciting from policy) at least one *baseline safe* price vector $p^{(0)}$ such that $c_j(p^{(0)}) \le 0$ for all j; for example, a uniform price that is known to satisfy parity by construction, or a historically used pricing scheme that has passed compliance review. If $p^{(0)}$ is known safe, we can seed the GP with an initial observation at $p^{(0)}$ and, if needed, impose the additional rule that $p^{(0)} \in \mathcal{A}_t^{safe}$ for all t (e.g., by never shrinking the safe set in a way that excludes it). When safety is not known deterministically, a weaker but still disciplined alternative is to begin with a short calibration phase in which we restrict attention to a conservative subset $\underline{\mathcal{P}} \subset [p_\ell, p_h]^G$ that is believed to be safe under minimal assumptions, using those data to form the first nontrivial $\mathcal{A}_t^{safe}$.

Second, feasibility must be reconciled with volatility limits. Let $\mathcal{V}_t$ denote the operational constraint set implied by the seller's allowed price adjustments at time t. For example, under a total-variation budget we may use a local TV ball

$$\mathcal{V}_t^{TV} \;=\; \Big\{p \in [p_\ell, p_h]^G : \; \|p - p_{t-1}\|_1 \le v_t\Big\},$$

with $\sum_{t=2}^{T} v_t \leq V$, or under a hard switch budget we may enforce piecewise-constant pricing by restricting changes to block boundaries. The action actually available at time t is then the intersection $\mathcal{A}_t^{safe} \cap \mathcal{V}_t$. This intersection captures a policy-relevant tension: strict governance shrinks the feasible set *across groups*, while volatility limits shrink it *over time.* The algorithmic choice rule (stated earlier) simply maximizes the optimistic revenue score over this doubly restricted region.

Safe sets expand endogenously with learning. Because $\sigma_{j,t}^c(p)$ shrinks around queried actions (at a rate governed by Γ_T), the conservative slack $\sqrt{\beta_{t+1}}\sigma_{j,t}^c(p)$ contracts over time, and additional price vectors satisfy $U_{j,t}^c(p) \leq 0$. Economically, compliance becomes less distortive as the platform accumulates evidence about how prices map into parity gaps or outcome disparities: rules that are initially binding due to uncertainty become binding only where they are substantively restrictive. This expansion is not monotone for arbitrary β_t, but with nondecreasing β_t and standard GP updating, the dominant force is variance reduction, so the seller typically experiences a growing menu of certified actions.

We emphasize a limitation: conservative feasibility can be overly restrictive when constraint feedback is noisy or indirect (e.g., outcome fairness inferred from conversion estimates). In such cases, $\mathcal{A}_t^{safe}$ may remain small for long horizons, motivating either richer data collection (to reduce σ^c) or a controlled relaxation via the primal–dual variant in which temporary violations are penalized rather than forbidden.

7 5.2. Safe action set construction (conservative constraints) and safe set expansion.

5.3. Primal–dual extension when safety constraints are soft (bounded violation). Conservative certification is attractive when governance rules must hold *period-by-period.* In many deployments, however, regulators and internal risk teams permit limited and auditable noncompliance—for instance, a small, transient outcome disparity while a new market is explored, or occasional parity deviations due to operational frictions—provided that violations are controlled and converge to zero on average. When such *soft* constraints are acceptable, we can replace the hard safe-set restriction with a primal–dual mechanism that learns the shadow prices of governance.

We introduce the (unknown) constraint vector $c(p) = (c_1(p), \ldots, c_J(p))$ and define the Lagrangian objective

$$\mathcal{L}(p, \lambda) \;=\; F(p) - \sum_{j=1}^{J} \lambda_j \, c_j(p), \qquad \lambda \in \mathbb{R}_+^J.$$

Economically, the dual variables λ_j act as endogenously chosen penalties: if constraint j has been repeatedly violated, its multiplier rises, making future actions that risk further violation less attractive even when they promise high revenue. This is precisely the "governance cost" interpretation that is often used in compliance discussions: instead of banning actions outright, we price their expected governance externality.

Operationally, we combine GP optimism for revenue with a conservative (or simply point-estimated) assessment of constraints inside the Lagrangian. A convenient choice rule is

$$p_t \in \arg\max_{p \in \mathcal{V}_t} \Big(U^F_{t-1}(p) - \sum_{j=1}^{J} \lambda_{j,t}\, U^c_{j,t-1}(p) \Big),$$

where $\mathcal{V}_t$ encodes the volatility restriction (block-constant pricing under a switch budget, or a local TV ball under a variation budget). This rule is intentionally modular: the platform continues to explore optimistically in revenue space, but exploration that is predicted to increase governance risk is automatically discounted by the current multipliers. In settings where U^c is overly pessimistic, one may instead use posterior means μ^c (or a one-sided bound) to avoid paralyzing the policy; the analysis then tracks the additional estimation error as an additive term.

After posting p_t and observing noisy constraint feedback, we update λ via projected gradient ascent (or mirror descent) on the dual:

$$\lambda_{j,t+1} = \Big[\lambda_{j,t} + \eta_t\, \widehat{c}_{j,t}(p_t) \Big]_+, \qquad j = 1, \dots, J,$$

where $\widehat{c}_{j,t}(p_t)$ is the observed proxy for $c_j(p_t)$ (e.g., an estimated parity gap or outcome disparity), $\eta_t > 0$ is a stepsize, and $[\cdot]_+$ denotes projection onto $\mathbb{R}_+$. The mechanism is transparent: if a constraint is satisfied (negative feedback), its multiplier drifts downward; if it is violated, the multiplier increases and raises the future implicit cost of similar actions.

Under standard conditions familiar from online convex optimization—in particular, a Slater-type condition ensuring that some action in $\mathcal{V}_t$ achieves strict feasibility with margin, boundedness of $c_j(p)$, and convex (or convex-surrogate) constraints—this primal–dual scheme yields two policy-relevant guarantees. First, regret relative to the best feasible comparator in the same volatility class remains sublinear, because the optimistic term U^F drives exploration while the dual variables prevent persistent governance drift. Second, cumulative violation

$$\mathrm{Viol}_T = \sum_{t=1}^{T} \sum_{j=1}^{J} [c_j(p_t)]_+$$

is $o(T)$, so that average violation vanishes even though occasional violations are allowed. Intuitively, the platform may "borrow" governance slack early to learn, but must "repay" it as multipliers rise; the long-run outcome is disciplined compliance without requiring an initially large certified action set.

We stress two limitations. If constraints are highly nonconvex in prices (as can occur with discontinuous parity rules or complex outcome metrics), primal–dual updates may stabilize only around local solutions; in such cases, convex relaxations or linearized constraints are often necessary. Moreover, if constraint feedback is severely delayed or biased (e.g., fairness inferred from sparse conversion data), the dual variables can react slowly, and stepsizes must be tuned conservatively. These caveats reflect a broader point: soft governance is most credible when measurement pipelines and auditing processes are strong enough that violations are not only bounded in theory but also observable and remediable in practice.

8 5.3. Primal–dual extension when safety constraints are soft (bounded violation).

5.4. Volatility control: (i) batching for hard switch budgets; (ii) regularized selection for switching costs/TV constraints. In many pricing deployments, the binding constraint is not only governance but also *operational volatility*: frequent price adjustments create engineering overhead, complicate customer communication, and can themselves trigger fairness scrutiny when groups observe rapidly changing differentials. We therefore treat volatility as a first-class constraint on the learning dynamics, rather than as an afterthought imposed by manual throttling. Conceptually, volatility control plays the same role as governance constraints: it restricts the action set over time and forces us to learn "under inertia."

(i) Batching to satisfy a hard switch budget. Under a hard cap S on the number of price changes, a simple and transparent design is to pre-commit to $B = S + 1$ blocks and keep prices constant within each block. Let $1 = \tau_1 < \tau_2 < \cdots < \tau_{B+1} = T + 1$ define a partition of periods into blocks $b = 1, \ldots, B$, where block b contains $t \in \{\tau_b, \ldots, \tau_{b+1} - 1\}$. We choose a block price vector $p^{(b)} \in [p_\ell, p_h]^G$ and set $p_t = p^{(b)}$ for all t in block b. This immediately guarantees $\#\{t \geq 2 : p_t \neq p_{t-1}\} \leq B - 1 = S$ deterministically, while still allowing the algorithm to explore across blocks.

From a learning perspective, batching converts the original T-period problem into a B-round GP bandit with *aggregated feedback.* If we define the block-average revenue observation

$$\bar{r}^{(b)} = \frac{1}{\tau_{b+1} - \tau_b} \sum_{t=\tau_b}^{\tau_{b+1}-1} \sum_{g \in \mathcal{G}} r_{g,t},$$

then $\bar{r}^{(b)}$ is a noisy sample of $F(p^{(b)})$ with reduced variance (sub-Gaussian scale shrinks with block length under conditional independence). Operationally, this aggregation is attractive: it matches common experimentation practice (hold a policy fixed long enough to measure it) and makes governance auditing easier because each block is a well-defined "pricing episode." The cost is adaptivity: mistakes persist for the duration of a block, so the regret bound scales with the effective horizon B, and block lengths must be chosen to balance measurement quality against responsiveness.

(ii) Regularized selection under switching costs or total variation (TV) budgets. When volatility is better modeled as a *cumulative* budget V (or as an explicit switching cost), we can encode it directly in the period-by-period choice rule. A convenient device is to maintain remaining variation V_{t-1} and restrict choices to a local neighborhood:

$$\mathcal{V}_t(V_{t-1}) \;=\; \Big\{p \in [p_\ell, p_h]^G :\; \|p - p_{t-1}\|_1 \le v_t,\; \sum_{s \le t} \|p_s - p_{s-1}\|_1 \le V\Big\},$$

where v_t can be a per-period cap that prevents abrupt jumps even when budget remains. We then select prices by solving a *regularized optimistic* problem, for example

$$p_t \in \arg\max_{p \in \mathcal{V}_t(V_{t-1})} \Big(U^F_{t-1}(p) \;-\; \rho_t \|p - p_{t-1}\|_1\Big),$$

or, when combined with governance penalties (hard-safe or primal–dual), by adding the corresponding constraint terms inside the same objective. Economically, ρ_t acts like an endogenous adjustment cost: it prices the operational disruption of changing group prices, and it prevents the algorithm from chasing small optimistic gains that are statistically fragile.

This regularization view is useful in practice because it admits continuous tradeoffs: rather than a brittle "change/no-change" rule, the algorithm can move gradually as posterior uncertainty shrinks. Computationally, the ℓ_1 term also encourages sparse adjustments (only a few groups move each period), which often matches how pricing teams implement rollouts.

Limitations and design guidance. Batching is robust and easily auditable, but it delays reaction to demand shifts and can amplify regret under nonstationarity. TV-regularized policies respond more smoothly, but require careful calibration of ρ_t (or v_t) to avoid getting stuck near an early baseline. In both cases, volatility control interacts with governance: stable pricing makes fairness metrics less noisy and easier to certify, but it also means that any initial fairness miscalibration can persist longer. Our recommendation is therefore pragmatic: use batching when operational change control is the primary constraint, and use TV regularization when gradual adaptation is feasible and monitoring pipelines are strong enough to detect drift quickly.

9 5.4. Volatility control: (i) batching for hard switch budgets; (ii) regularized selection for switching costs/TV constraints.

6. Theory. We now formalize the guarantees that justify governed learning under operational inertia. The technical core is standard GP concentration in RKHS combined with a restriction of the admissible action set induced by either (i) a hard switch budget or (ii) a total-variation (TV) budget. The main message is that volatility control changes *where* we are allowed to optimize and explore, but (under mild regularity) it does not change *how* posterior uncertainty accumulates: regret remains controlled by information gain, up to the effective horizon implied by the volatility constraint.

Confidence structure. Let $\mu_{t-1}^F(p)$ and $\sigma_{t-1}^F(p)$ denote the GP posterior mean and standard deviation for total revenue $F(p)$, and analogously $(\mu_{j,t-1}^c(p), \sigma_{j,t-1}^c(p))$ for each constraint $c_j(p)$. For a confidence parameter β_t chosen as in GP-UCB (scaling with Γ_t and $\log(1/\delta)$), define

$$U_{t-1}^F(p) = \mu_{t-1}^F(p) + \sqrt{\beta_t}\,\sigma_{t-1}^F(p), \quad U_{j,t-1}^c(p) = \mu_{j,t-1}^c(p) + \sqrt{\beta_t}\,\sigma_{j,t-1}^c(p).$$

Under the RKHS assumptions $\|f_g\|_{\mathcal{H}_k} \le B$, $\|c_j\|_{\mathcal{H}_k} \le B_c$, and σ-sub-Gaussian observation noise, we have with probability at least $1-\delta$ the simultaneous concentration event

$$\forall t,\ \forall p \in [p_\ell, p_h]^G: \quad |F(p) - \mu_{t-1}^F(p)| \le \sqrt{\beta_t}\,\sigma_{t-1}^F(p), \qquad |c_j(p) - \mu_{j,t-1}^c(p)| \le \sqrt{\beta_t}\,\sigma_{j,t-1}^c(p).$$

This event is the sole ingredient needed to obtain both safety (via conservative feasibility) and regret control (via optimism).

Safety under conservative feasibility. Define the safe set

$$\mathcal{A}_t^{safe} = \{p \in [p_\ell, p_h]^G : \ U_{j,t-1}^c(p) \le 0 \ \ \forall j\}.$$

Choosing $p_t \in \mathcal{A}_t^{safe}$ guarantees $c_j(p_t) \le 0$ for all t, j on the concentration event, because $c_j(p_t) \le U_{j,t-1}^c(p_t) \le 0$. In applications, we emphasize the practical implication: as long as the constraint feedback (possibly via proxies such as conversion-rate disparities) is calibrated so that the GP model is valid, governance auditing can be reduced to checking the algorithm's conservative bound U^c rather than the unknown c.

Regret with volatility classes. Let Π_S denote feasible price sequences with at most S switches and Π_V those with TV at most V. Consider the governed optimistic rule

$$p_t \in \arg\max_{p \in \mathcal{A}_t^{safe} \cap \mathcal{V}_t} U_{t-1}^F(p),$$

where $\mathcal{V}_t$ encodes either the batching restriction (hard-switch) or a local TV restriction. On the concentration event, instantaneous regret is bounded by the posterior width at the chosen point:

$$F(p_t^*) - F(p_t) \;\le\; U_{t-1}^F(p_t) - F(p_t) \;\le\; 2\sqrt{\beta_t}\,\sigma_{t-1}^F(p_t),$$

for any comparator p_t^* that remains feasible whenever p_t is chosen (in particular, for the best feasible policy in the same volatility class). Summing over t and applying the standard information-gain inequality yields

$$\mathrm{Reg}_T(\Pi) \;=\; \tilde{O}\big(\sqrt{T\,\Gamma_T}\big),$$

with Γ_T the maximum information gain under kernel k on $[p_\ell, p_h]^G$ (extended in the usual way to vector-valued observations when learning group revenues and constraints jointly).

Hard switches via batching. If we pre-commit to $B = S + 1$ blocks and use aggregated observations within each block, then the effective decision horizon becomes B. Under conditional independence, the block-average noise is sub-Gaussian with scale shrinking with block length, improving estimation within each chosen price. The resulting bound takes the form $\tilde{O}(\sqrt{B\,\Gamma_B})$ in block-time, which translates to a period-time regret that trades off infrequent updates against more reliable measurements.

TV budgets and switching costs. For TV control, we restrict to $\mathcal{V}_t = \{p : \|p - p_{t-1}\|_1 \le v_t\}$ or equivalently maximize a regularized optimistic objective $U_{t-1}^F(p) - \rho_t\|p - p_{t-1}\|_1$. The theory proceeds identically, except that $\mathcal{V}_t$ can reduce exploration early; accordingly, the comparator class must be restricted to sequences with comparable variation, matching the operational premise that rapid experimentation is infeasible.

Scope and limitations. These guarantees are only as strong as the modeling assumptions: severe nonstationarity, misspecified kernels, or biased constraint proxies can invalidate safety. In practice, we view the theory as a disciplined baseline: it clarifies which quantities must be monitored (posterior widths, safe-set non-emptiness, and variation consumption) and where additional robustness mechanisms (change-point tests, conservative priors, or primal–dual soft constraints) are needed.

9.1 6.1. High-probability feasibility (always-safe constraints) or sublinear cumulative constraint violation (primal–dual).

Our first governance objective is *ex ante* operational: at every period t, the posted price vector should satisfy the (unknown) constraints $c_j(p_t) \le 0$. The central idea is to convert an unobserved feasibility requirement into an *observable* one by acting only on the conservative region implied by the GP posterior. Concretely, whenever we enforce feasibility through the upper confidence bounds $U_{j,t-1}^c(\cdot)$, we are not “predicting” that a policy is safe; we are requiring that it is safe under the most adverse realization consistent with past data at confidence level $1 - \delta$.

Formally, let $\mathcal{E}$ denote the joint concentration event for the constraint models: for all t and all $p \in [p_\ell, p_h]^G$,

$$c_j(p) \;\le\; U^c_{j,t-1}(p) \qquad \forall j \in \{1, \dots, J\}.$$

By standard RKHS–GP concentration (with β_t chosen to account for a union bound over t and j), we have $\Pr(\mathcal{E}) \ge 1 - \delta$. On $\mathcal{E}$, feasibility becomes immediate: if the algorithm selects p_t such that $U^c_{j,t-1}(p_t) \le 0$ for every j, then $c_j(p_t) \le 0$ for every j. In other words, *always-safe* behavior is achieved not by learning constraints perfectly, but by never stepping outside a set that is provably feasible given current uncertainty. This yields a pathwise guarantee: with probability at least $1 - \delta$, there are *no* constraint violations at any time.

Two operational caveats are worth making explicit. First, the safe set must be non-empty at the times we need to act. In practice, we treat this as a design requirement: we initialize from a known compliant baseline price vector $p^{(0)}$ (often supplied by policy, prior auditing, or legacy pricing rules), seed the constraint GP with conservative prior mean, and restrict early optimization to a neighborhood of $p^{(0)}$ until posterior uncertainty shrinks. Second, safety depends on the fidelity of the constraint feedback channel. When constraints are computed from proxies (e.g., estimated conversion gaps used to audit outcome fairness), we require that the proxy error is either explicitly modeled as part of the observation noise or bounded so that the GP confidence band remains valid; otherwise, conservative feasibility can provide false reassurance.

When strict per-period safety is too conservative—for example, because early uncertainty makes $\mathcal{A}^{safe}_t$ small, or because the organization is willing to tolerate rare, small violations in exchange for faster learning—we can instead target *sublinear cumulative violation.* A convenient approach is primal–dual learning on a Lagrangian relaxation. Let $\lambda_t \in \mathbb{R}^J_+$ be dual variables (interpretable as shadow prices of governance). At time t, we choose a price vector by solving a penalized optimistic problem such as

$$p_t \in \arg\max_{p \in [p_\ell, p_h]^G \cap \mathcal{V}_t} \Big(U^F_{t-1}(p) \;-\; \sum_{j=1}^{J} \lambda_{j,t}\, U^c_{j,t-1}(p) \Big),$$

and then update λ_t by projected subgradient ascent using realized (or estimated) constraint feedback, e.g.

$$\lambda_{j,t+1} \;=\; \Big[\lambda_{j,t} + \eta_t\, c_j(p_t) \Big]_+ .$$

Under standard conditions used in online convex optimization—most importantly, a Slater-type condition ensuring a strictly feasible point exists, and bounded gradients for the chosen constraint surrogates—one obtains a tradeoff: regret against the best feasible comparator remains sublinear while $\mathrm{Viol}_T = \sum_{t=1}^{T}\sum_{j=1}^{J}[c_j(p_t)]_+$ satisfies $\mathrm{Viol}_T = o(T)$. Economically, $\lambda_{j,t}$ learns the "price" of governance: if the algorithm repeatedly drifts toward infeasible regions, dual penalties grow and redirect future choices back toward compliance.

We view the always-safe and primal–dual modes as complementary. The former is appropriate when governance is a hard requirement (regulated parity, contractual obligations), whereas the latter matches settings where governance is enforced through audits, remediation, or expected-penalty regimes and where the organization prefers a smooth learning–compliance frontier rather than a hard feasibility barrier.

9.2 6.2. Regret versus the best feasible policy under volatility constraints; dependence on the number of groups and constraints.

We now turn from feasibility to performance: how much revenue we forego, relative to the best *governed* pricing policy that respects the same operational limits on price movement. Given a comparator class Π (e.g., a hard-switch class Π_S or a total-variation class), we measure learning performance by cumulative regret

$$\mathrm{Reg}_T(\Pi) := \sum_{t=1}^{T}\big(F(p_t^\star)-F(p_t)\big), \qquad \{p_t^\star\} \in \arg\max_{\{p_t\}\in\Pi}\sum_{t=1}^{T}F(p_t) \text{ s.t. } c_j(p_t)\le 0\,\forall j,t,$$

so that the benchmark internalizes both governance and volatility.

The key observation is that conservative governance changes *where* we may search but not *how* uncertainty translates into regret. On the joint concentration event for the revenue model, we have $F(p) \le U_{t-1}^F(p)$ for all p, hence for any feasible comparator action $p_t^\star$ admissible at time t,

$$F(p_t^\star) - F(p_t) \;\le\; U_{t-1}^F(p_t^\star) - F(p_t) \;\le\; U_{t-1}^F(p_t) - F(p_t),$$

where the final inequality uses that the algorithm maximizes U_{t-1}^F over the admissible set $\mathcal{A}_t^{safe} \cap \mathcal{V}_t$. Standard GP-UCB algebra then bounds instantaneous regret by the posterior width at the chosen point, typically

$$U_{t-1}^F(p_t) - F(p_t) \;\le\; 2\sqrt{\beta_t}\,\sigma_{t-1}^F(p_t),$$

and summing these widths over time yields

$$\mathrm{Reg}_T(\Pi) \;=\; \tilde{O}\Big(\sqrt{T\,\Gamma_T}\Big),$$

up to constants depending on (B,σ) and logarithmic factors in $(1/\delta)$. The role of governance enters through the requirement that the comparator sequence remains feasible and, under always-safe operation, that it lies within the safe region induced by the confidence bounds (or that the safe set is rich enough to contain an optimal governed action). When we instead use a primal–dual relaxation, the same uncertainty accounting controls the *objective* regret, while the dual analysis ensures $\mathrm{Viol}_T = o(T)$ under the usual Slater-type condition.

Volatility constraints affect regret through an effective reduction in the decision horizon. Under a hard switch budget S, a canonical construction is batching: partition $\{1, \dots, T\}$ into $B = S + 1$ blocks and restrict p_t to be constant within each block. We then learn in B decision rounds with aggregated (lower-variance) feedback, obtaining GP-UCB regret $\tilde{O}(\sqrt{B\,\Gamma_B})$ at the block level, which translates into period regret of order $\tilde{O}(\sqrt{T^2\Gamma_B/B})$ after accounting for block lengths. Under a total-variation budget V, one can instead restrict $\mathcal{V}_t$ to an ℓ_1 ball around p_{t-1}, which typically yields similar bounds with additional terms reflecting how tightly $\mathcal{V}_t$ restricts exploration.

Finally, we highlight how the number of groups G and constraints J enter. Statistically, G raises the complexity of the action space $[p_\ell, p_h]^G$, and thus the information gain Γ_T; for many kernels, Γ_T grows at least polylogarithmically in T with exponents that worsen in dimension, making large-G problems intrinsically harder without additional structure. Governance can partially offset this by *coupling* prices across groups (e.g., parity constraints reduce dispersion and can lower the effective dimension). The constraint count J mainly appears through confidence calibration—β_t must absorb a union bound over (t, j), producing logarithmic dependence on J—and through computation, since each candidate p must be screened against J conservative bounds. In practice, scaling to many groups and constraints hinges on exploiting shared structure (e.g., low-rank kernels, hierarchical group models, or constraint decompositions) and on reliable numerical maximization of the constrained acquisition problem.

9.3 6.3. Discussion of kernel misspecification and practical hyperparameter tuning.

Our analysis has treated the kernel k (and the associated RKHS radius bounds B, B_c) as known, which is analytically convenient but empirically demanding. In practice, the principal failure mode is *kernel misspecification*: the true mean revenue and constraint functions may not lie in $\mathcal{H}_k$, or they may satisfy the smoothness assumptions only approximately. This matters twice. First, the usual GP concentration inequalities that underwrite $F(p) \leq U_t^F(p)$ and $c_j(p) \leq U_{j,t}^c(p)$ can become anti-conservative, in which case both regret bounds and (more importantly) period-by-period safety may fail. Second, an ill-tuned kernel can slow learning by allocating posterior variance in the wrong regions of price space, effectively shrinking the useful safe set $\mathcal{A}_t^{safe}$.

A useful way to interpret misspecification is as an approximation error decomposition. Suppose the true function admits a best RKHS approximation $f_g^\dagger \in \mathcal{H}_k$ with $\|f_g^\dagger\|_{\mathcal{H}_k} \leq B$, and write $f_g = f_g^\dagger + \xi_g$, where ξ_g is a residual. Then the GP-UCB regret logic typically persists but acquires an additive bias term proportional to the cumulative impact of ξ_g. Heuristically, one expects bounds of the form

$$\mathrm{Reg}_T(\Pi) \ \lesssim\ \tilde{O}(\sqrt{T\Gamma_T}) \ +\ T \cdot \sup_{p\in[p_\ell,p_h]^G} \Big|\sum_g \xi_g(p(g))\Big|,$$

and an analogous degradation in constraint guarantees, where a residual ξ_j^c can be interpreted as an unmodeled constraint drift. While such statements are necessarily informal without additional structure, they emphasize an operational lesson: when governance is binding, it is typically more important to be conservative about constraint modeling than to be perfectly efficient about revenue modeling.

Kernel choice is our first lever for robustness. When little is known about smoothness, we prefer kernels that do not hard-code excessive differentiability (e.g., Matérn kernels with modest smoothness) over very smooth squared-exponential kernels. When G is large, structure helps: additive kernels across groups, low-rank multi-task kernels, or hierarchical priors that share hyperparameters can substantially reduce sample complexity by pooling information, while still allowing group idiosyncrasies. For constraints that encode parity or outcome fairness, the relevant function may be closer to a difference of two smooth surfaces (e.g., $\mu_g(p) - \mu_{g'}(p')$), suggesting kernels that respect such algebraic structure.

Hyperparameter tuning must be handled with particular care because our data are adaptively collected. A common practical approach is empirical Bayes: at each update, re-fit kernel hyperparameters (lengthscales, output variance, noise scale) by maximizing the GP marginal likelihood (possibly with priors/regularization), and then recompute posteriors. This often performs well, but it breaks the strict assumptions behind fixed-k concentration unless one inflates confidence widths. A simple safeguard is to treat hyperparameter learning as part of uncertainty: choose β_t and the norm bounds (B, B_c) conservatively enough that, for a range of plausible hyperparameters, the resulting confidence bands remain valid. In governed applications, we often recommend a "safety-first" calibration: err on the side of larger σ (noise) and larger β_t, accepting slower learning in exchange for fewer governance surprises.

Two additional engineering details are worth emphasizing. First, one should seed the procedure with at least one *certifiably feasible* baseline price vector p^{safe} (e.g., a uniform price satisfying parity by construction, or a regulator-approved menu). This ensures $\mathcal{A}_t^{safe}$ is non-empty even if early hyperparameter estimates are unstable. Second, constraint feedback is frequently indirect (conversion proxies, survey-based parity metrics, delayed chargeback rates), so the observation model for c_j may be heteroscedastic and non-Gaussian. In that case, a Gaussian likelihood is a pragmatic approximation, but we can further stabilize safety by using robust regression (e.g., Student-t likelihoods) or by placing conservative caps on effective signal-to-noise when forming $U^c_{j,t}$.

Finally, we note a limitation that interacts sharply with volatility budgets. If S (or V) is very tight, the algorithm has few opportunities to correct a poor kernel fit, and hyperparameter learning itself may "overfit" to a small set of queried prices. Practically, batching can be paired with periodic re-estimation of hyperparameters at block boundaries, using aggregated data to reduce variance. More broadly, when governance risk is material, the economically relevant objective is not the tightest theoretical regret rate, but a stable operating procedure whose safety conclusions remain credible under mild misspecification; conservative tuning is often the right institutional choice.

10 6.3. Discussion of kernel misspecification and practical hyperparameter tuning.

Our theoretical development conditions on a known kernel k and known complexity radii (B, B_c). This is a deliberate simplification: in deployed pricing systems the kernel is itself a modeling choice, and its hyperparameters (lengthscales, output variance, observation noise) are estimated from adaptively collected data. The central practical concern is *misspecification.* If the true mean revenue f_g or a constraint surface c_j is rougher, more non-stationary, or structurally different than $\mathcal{H}_k$ permits, then the GP posterior can become systematically overconfident. In governed settings this is not merely a statistical nuisance: anti-conservative confidence bands can cause the safe set $\mathcal{A}_t^{safe}$ to contain points that are in fact infeasible, undermining period-by-period guarantees precisely when the institution cares most about them.

A useful lens is an approximation-error decomposition. Let $f_g^\dagger \in \mathcal{H}_k$ denote the best approximation to f_g under the chosen kernel and norm budget, and write $f_g = f_g^\dagger + \xi_g$, with residual ξ_g capturing the misspecification component. Then the classical optimistic-regret argument still bounds the part of regret attributable to uncertainty about $f_g^\dagger$, but it cannot remove the bias introduced by ξ_g. Heuristically, for a feasible comparator class Π,

$$\mathrm{Reg}_T(\Pi) \;\lesssim\; \tilde{O}\Big(\sqrt{T\Gamma_T}\Big) \;+\; \sum_{t=1}^{T} \sup_{p\in[p_\ell,p_h]^G} \Big|\sum_{g\in\mathcal{G}} \xi_g(p(g))\Big|,$$

and an analogous term appears for each constraint via $c_j = c_j^\dagger + \xi_j^c$. The operational message is that misspecification is "paid" in a linear-in-T way unless the residuals are uniformly small. This is why, when governance constraints bind, we generally prefer to spend modeling effort (and conservatism) on constraints rather than on marginal improvements in revenue fit.

Kernel choice is our first robustness lever. When smoothness is uncertain, overly rigid priors (e.g., squared-exponential kernels with long lengthscales) can be fragile: they interpolate aggressively and can underestimate uncertainty away from observed prices. Matérn families with moderate smoothness, or mixtures that allow multiple lengthscales, tend to degrade more gracefully. When G is large, the challenge is not only statistical but computational: multi-task structure (e.g., a separable kernel $k((g,p),(g',p')) = k_G(g,g')k_P(p,p')$) or additive decompositions can pool information across groups while still permitting group-specific idiosyncrasies. For parity- or disparity-type constraints, kernels that respect the algebraic form (differences across groups, monotone transformations of demand) can materially improve sample efficiency because the constraint learns from structured comparisons rather than from raw levels.

Hyperparameter tuning must be handled carefully because our data are collected under feedback: the algorithm's past posteriors shape future price queries, which in turn shape the likelihood surface. Empirical Bayes (reestimating hyperparameters by marginal likelihood maximization at each step) is often effective, but it breaks the fixed-kernel concentration logic unless we hedge. In practice, we recommend treating hyperparameter uncertainty as part of the safety margin: inflate β_t, and choose (B, B_c) and the noise scale conservatively so that confidence bands remain credible across a plausible hyperparameter set. A complementary approach is to update hyperparameters only at coarse timescales (e.g., block boundaries under batching), which reduces adaptivity and stabilizes estimation.

Two implementation safeguards are especially valuable. First, the procedure should be initialized with a *certifiably feasible* price vector p^{safe} so that $\mathcal{A}_t^{safe}$ is non-empty even under unstable early fits. Second, constraint feedback is often indirect and heteroscedastic (conversion proxies, audits, delayed complaints). If Gaussian observation models are used as approximations, one should cap effective signal-to-noise or adopt robust likelihoods (e.g., Student-t) when forming $U_{j,t}^c$, since outliers in constraint measurements can otherwise cause spurious "permission" to explore.

Finally, volatility budgets interact sharply with misspecification. If S (or V) is tight, the platform has few opportunities to recover from an early, poorly tuned kernel: exploratory corrections are expensive, and hyperparameter learning can overfit to a narrow region of prices. In such environments, conservative kernels, slow hyperparameter adaptation, and regulator-approved safe baselines are not merely technical choices; they are institutional risk controls that trade a small amount of revenue optimality for substantially greater credibility of governance compliance.

10.1 Synthetic environments (nonstationary, non-monotone) with fairness ground truth.

We begin our experimental analysis in synthetic environments where the seller faces group-dependent, nonlinear demand responses, yet we retain full knowledge of the ground-truth revenue and governance constraints. The purpose is twofold. First, we can compute credible comparators (including volatility-constrained ones) and therefore report meaningful regret and violation statistics. Second, we can deliberately introduce the kinds of nonstationarity and non-monotonicity that are difficult to rule out in practice, thereby stress-testing whether governed exploration remains disciplined when the model class is only an approximation.

Groups, prices, and non-monotone demand. We fix $G \in \{3, 5, 10\}$ groups and a compact price interval $[p_\ell, p_h]$, and generate for each group a mean demand curve that is *not* globally decreasing in price. Concretely, we construct

$$\mu_{g,t}(p) \;=\; \Big(a^{(1)}_{g,t}\exp\big(-\tfrac{(p-b^{(1)}_{g,t})^2}{2(s^{(1)}_g)^2}\big) \;+\; a^{(2)}_{g,t}\exp\big(-\tfrac{(p-b^{(2)}_{g,t})^2}{2(s^{(2)}_g)^2}\big)\Big)_+,$$

so that each group has two “preference modes” (e.g., bargain-seekers and convenience buyers) and the implied revenue surface $f_{g,t}(p) = p\,\mu_{g,t}(p)$ can exhibit multiple local maxima. We cap $\mu_{g,t}(p)$ if needed to ensure bounded rewards. This construction yields a controlled violation of the canonical monotone-demand assumption without resorting to adversarial worst cases.

Nonstationarity mechanisms. To model drift and regime changes, we let the bump locations and amplitudes vary over time:

$$b^{(i)}_{g,t} = b^{(i)}_g + \rho^{(i)}_g \sin\Big(\frac{2\pi t}{\tau}\Big) + \kappa^{(i)}_g \mathbf{1}\{t \geq t_0\}, \qquad a^{(i)}_{g,t} = a^{(i)}_g\big(1 + \nu \sin(\tfrac{2\pi t}{\tau_a})\big),$$

where the sinusoidal terms represent seasonal fluctuations and the indicator term represents a one-time shock (e.g., a competitor entry) at t_0. By varying (ρ, κ, ν), we sweep from nearly stationary environments (where RKHS priors are plausible) to strongly nonstationary ones (where any fixed-kernel model is inevitably misspecified). Importantly, the governance constraints below are defined on $\mu_{g,t}$ and therefore inherit the same temporal structure.

Observation model and constraint feedback. At each period t, we generate realized revenue

$$r_{g,t} \;=\; f_{g,t}\big(p_t(g)\big) + \epsilon_{g,t},$$

with $\epsilon_{g,t}$ drawn from a mean-zero noise distribution calibrated to be sub-Gaussian, and optionally heteroscedastic in price (larger near p_ℓ and p_h)

to mimic thin data regimes. For outcome-based constraints, we separately generate a noisy proxy for expected purchase outcomes, e.g.,

$$\widehat{\mu}_{g,t}\big(p_t(g)\big) \;=\; \mu_{g,t}\big(p_t(g)\big) + \zeta_{g,t},$$

which we interpret as an estimable conversion or audit signal; this allows us to evaluate algorithms that maintain a distinct GP (or surrogate) for $c_j(\cdot)$.

Fairness ground truth and feasibility. We include (i) price parity constraints $|p(g) - p(g')| - \Delta \le 0$ and (ii) outcome-fairness constraints defined directly from the synthetic demand,

$$c^{out}_{g,g',t}(p) \;=\; \big|\mu_{g,t}(p(g)) - \mu_{g',t}(p(g'))\big| - \varepsilon \;\le\; 0,$$

with ε chosen so that the feasible set is non-empty but nontrivial. We also specify a certifiably feasible baseline p^{safe} (typically a common midrange price) and verify $c_{j,t}(p^{safe}) \le 0$ for all t by construction, ensuring that safety-oriented methods are well-posed from the first round.

Evaluation protocol and comparators. For each environment instance we run multiple random seeds and report: cumulative regret against the best feasible comparator in the same volatility class (hard-switch budget S via batching, or TV budget V); cumulative and peak violations $\sum_{t,j}[c_{j,t}(p_t)]_+$ and $\max_{t,j}[c_{j,t}(p_t)]_+$; and realized volatility (switch count and total variation) as a diagnostic of implementation fidelity. Because the action space is continuous, we approximate the comparator by discretizing prices on a fine grid and solving the resulting volatility-constrained planning problem (a shortest-path or dynamic program under S or V), which yields an explicit, auditable benchmark. This synthetic suite therefore isolates the fundamental question of governed learning: how quickly can we approach the best *feasible and operationally stable* policy when the true environment is smooth only in parts and changes over time.

11 7.1. Synthetic environments (nonstationary, nonmonotone) with fairness ground truth.

We construct a family of synthetic testbeds that are deliberately rich along the dimensions that matter for governed learning: (i) group heterogeneity, (ii) non-concavity in the per-group revenue landscape, and (iii) time variation that ranges from mild drift to abrupt regime change. The central benefit of working synthetically is that we can treat both the revenue objective and the governance constraints as *ground truth* objects, so that regret, feasibility, and operational stability can be audited without ambiguity.

For each instance we draw $G \in \{3, 5, 10\}$ groups and restrict prices to a compact interval $[p_\ell, p_h]$. Group-level demand is generated from a two-mode specification,

$$\mu_{g,t}(p) = \Big(a_{g,t}^{(1)} \exp\Big(-\frac{(p - b_{g,t}^{(1)})^2}{2(s_g^{(1)})^2}\Big) + a_{g,t}^{(2)} \exp\Big(-\frac{(p - b_{g,t}^{(2)})^2}{2(s_g^{(2)})^2}\Big)\Big)_+, \qquad f_{g,t}(p) = p\,\mu_{g,t}(p),$$

where the truncation $(\cdot)_+$ avoids negative demand when we add noise or shocks. Intuitively, the two bumps represent two latent buyer segments within each protected class (e.g., a price-sensitive segment and a convenience segment). Because the bumps can overlap, $f_{g,t}$ is typically *non-unimodal*, and its maximizer can move non-monotonically as the environment drifts. To avoid degenerate reward scales across groups, we cap $\mu_{g,t}(p)$ at a fixed $\overline{\mu}$, which also ensures bounded revenues on $[p_\ell, p_h]$.

Nonstationarity enters by allowing both locations and amplitudes to evolve:

$$b_{g,t}^{(i)} = b_g^{(i)} + \rho_g^{(i)} \sin\Big(\frac{2\pi t}{\tau_b}\Big) + \kappa_g^{(i)}\,\mathbf{1}\{t \geq t_0\}, \qquad a_{g,t}^{(i)} = a_g^{(i)}\Big(1 + \nu \sin\Big(\frac{2\pi t}{\tau_a}\Big)\Big),$$

where ρ controls smooth seasonal drift and κ induces an interpretable one-time shock (e.g., entry, policy change, or a measurement pipeline update) at time t_0. By tuning (ρ, κ, ν) we obtain a spectrum from near-stationary instances—where a fixed-kernel GP prior is a reasonable approximation—to instances where any time-invariant model is misspecified, so that safety mechanisms must operate under model error rather than purely statistical uncertainty.

Given posted prices $p_t \in [p_\ell, p_h]^G$, we generate realized revenue observations via

$$r_{g,t} = f_{g,t}\big(p_t(g)\big) + \epsilon_{g,t},$$

with $\epsilon_{g,t}$ mean-zero and calibrated to be σ-sub-Gaussian; in some instances we let the variance increase near the boundaries p_ℓ and p_h to mimic thin-data regions. For outcome-based governance, we also generate a distinct noisy proxy for purchase outcomes,

$$\widehat{\mu}_{g,t}\big(p_t(g)\big) = \mu_{g,t}\big(p_t(g)\big) + \zeta_{g,t},$$

which we interpret as an auditable conversion estimate, allowing separate learning for revenue and for constraint-relevant quantities.

We impose two canonical constraint families. Price parity is enforced through $|p(g) - p(g')| - \Delta \leq 0$. Outcome fairness is defined directly on the demand primitives,

$$c_{g,g',t}^{out}(p) = \big|\mu_{g,t}(p(g)) - \mu_{g',t}(p(g'))\big| - \varepsilon \leq 0,$$

with ε chosen so that the feasible set is non-empty yet meaningfully restrictive. To ensure well-posed safe learning from $t = 1$, we explicitly construct a certifiably feasible baseline p^{safe} (typically a common midrange price) and verify by construction that $c_{j,t}(p^{safe}) \leq 0$ for all t and all imposed constraints.

Evaluation proceeds by averaging over multiple random seeds per instance and reporting (i) cumulative regret relative to the best feasible comparator within the same volatility class (hard switch budget S via batching, or total variation budget V), (ii) cumulative and peak constraint violations $\sum_{t,j}[c_{j,t}(p_t)]_+$ and $\max_{t,j}[c_{j,t}(p_t)]_+$, and (iii) realized volatility (switch counts and total variation). Because the action space is continuous, we approximate the comparator by discretizing $[p_\ell, p_h]$ on a fine grid and solving the resulting constrained planning problem (a shortest-path dynamic program under S, or a knapsack-like recursion under V). This delivers an explicit benchmark that is both strong and auditable, clarifying the extent to which governed learning can track the best *feasible and operationally stable* pricing policy even when the environment is nonlinear and evolving.

11.1 7.2. Semi-synthetic marketplace data: offline replay + safe online simulation.

Purely synthetic instances give us full control, but they may understate the institutional frictions that motivate governed learning in practice (missing covariates, selection in who sees which prices, and noisy fairness measurement). We therefore complement the synthetic testbeds with a semi-synthetic construction that starts from real marketplace logs and then layers a controlled, auditable simulation on top. The goal is to retain the empirical distribution of contexts and group composition while still giving ourselves a *ground truth* environment in which regret and constraint violations can be computed for counterfactual pricing rules.

We begin with an offline replay protocol. The logged dataset consists of tuples

$$\Big\{(x_t,\ g,\ p_t^{\log}(g),\ r_{g,t}^{\log},\ \widehat{\mu}_{g,t}^{\log})\Big\}_{t=1}^{T},$$

where $p_t^{\log}(g)$ is the historically posted group price, $r_{g,t}^{\log}$ is realized revenue, and $\widehat{\mu}_{g,t}^{\log}$ is an auditable proxy for purchase outcomes (e.g., estimated conversion). When the logging policy randomizes prices (even mildly), we exploit known or estimable propensities $\pi_t^{\log}(p \mid x_t, g)$ to evaluate candidate policies π via weighted replay. Concretely, for any policy that selects $p_t(g) = \pi(x_t, g)$ (possibly with volatility restrictions encoded in π), an inverse-propensity estimator of mean revenue is

$$\widehat{R}(\pi) = \frac{1}{T}\sum_{t=1}^{T}\sum_{g\in\mathcal{G}} \frac{\mathbf{1}\{p_t^{\log}(g) = \pi(x_t, g)\}}{\pi_t^{\log}(p_t^{\log}(g) \mid x_t, g)}\, r_{g,t}^{\log},$$

and similarly for constraint-relevant outcomes by replacing $r_{g,t}^{\log}$ with $\widehat{\mu}_{g,t}^{\log}$ inside the constraint definition. Because exact matching is brittle in continuous price spaces, we use a discretized price grid (or a small set of operational "price buckets") and treat policies as mapping into that finite action set; we also report stabilized and clipped weights to control variance. To reduce bias from limited overlap, we restrict attention to policy classes that do not stray far from the support of the logging policy—a restriction that is economically natural when governance requires incremental changes and audit trails.

Offline replay alone cannot provide the full suite of governed-learning diagnostics we want, because constraint satisfaction under π is only partially observed: we see outcomes at the logged prices, not at the counterfactual prices. We therefore fit flexible response models $\widehat{f}_g(p, x)$ and $\widehat{\mu}_g(p, x)$ (using nonparametric regressors or GPs on (p, x) with group-specific components) and combine them with replay through doubly robust scores. This produces policy-value and fairness estimators that remain consistent if either the propensity model or the outcome model is correctly specified, while retaining an explicit decomposition into statistical uncertainty (finite data) versus model uncertainty (misspecification). In reporting, we treat this step as an *audit primitive*: it makes clear which part of the evaluation is "as observed" and which part is model-imputed.

The second stage is a safe online simulation that is anchored in the replay data but restores a controlled ground truth. We construct a simulator in which the context sequence $\{x_t\}$ is taken directly from the logs (or resampled in blocks to preserve seasonality), while the conditional mean outcomes are given by a calibrated structural proxy,

$$\mu_{g,t}(p) = \widehat{\mu}_g(p, x_t) + \eta_{g,t}^{\mu}(p), \qquad f_{g,t}(p) = p\,\mu_{g,t}(p),$$

where $\eta_{g,t}^{\mu}(p)$ is a mean-zero residual process fitted to match empirical dispersion and, when desired, engineered to include regime shifts at known dates (e.g., policy changes) to stress-test safety under nonstationarity. Revenues are then generated as $r_{g,t} = f_{g,t}(p_t(g)) + \epsilon_{g,t}$, and governance constraints are computed from the simulator's $\mu_{g,t}$ so that violations are objectively measurable at every t. We also impose an explicit certified baseline p^{safe} (often a uniform price) and, by construction, ensure $c_j(p^{safe}) \leq 0$ for all simulated periods, so that safe exploration is feasible from $t = 1$.

This semi-synthetic pipeline reflects how governed pricing is deployed in practice: offline evidence is necessary but insufficient, and any credible evaluation must separate (i) what can be justified from historical exposure, from (ii) what is a model-based extrapolation that should be treated conservatively. Its main limitation is the usual one: if the logs lack support for certain price regions or if unobserved confounding drives both price assignment and demand, then neither replay nor simulation can fully identify counterfactual effects. We therefore treat the semi-synthetic results as complementary to the fully synthetic benchmarks, not as a substitute for them.

11.2 7.3. Metrics: revenue, regret proxies, fairness violations, volatility, stability, and auditability.

Our evaluation metrics mirror the objective and constraint language of the model while remaining implementable in offline replay and in semi-synthetic online simulation. We report metrics at the horizon level (T) and, when informative, as trajectories over t.

Revenue and uplift. The primary performance metric is realized cumulative revenue

$$\mathrm{Rev}_T \;=\; \sum_{t=1}^{T}\sum_{g\in\mathcal{G}} r_{g,t}, \qquad r_{g,t} = p_t(g)D_{g,t}(p_t(g)).$$

Because governance is often justified relative to an operational baseline, we also report *uplift* versus a certified baseline policy (typically a uniform or historically used price vector p^{safe}):

$$\mathrm{Uplift}_T \;=\; \frac{\mathrm{Rev}_T - \mathrm{Rev}_T^{safe}}{\mathrm{Rev}_T^{safe}},$$

which is interpretable even when the true optimum is unknown. In nonstationary semi-synthetic environments we additionally report block-level revenue to reveal whether performance is driven by early exploration or by sustained improvements under volatility restrictions.

Regret and regret proxies. When a ground-truth simulator is available (fully synthetic or calibrated semi-synthetic), we compute pseudo-regret against the *best feasible comparator* in the relevant volatility class:

$$\mathrm{Reg}_T(\Pi) \;=\; \sum_{t=1}^{T}\Big(F(p_t^*) - F(p_t)\Big), \qquad \{p_t^*\} \in \arg\max_{\{p_t\}\in\Pi}\sum_{t=1}^{T} F(p_t) \;\text{ s.t. }\; c_j(p_t) \le 0.$$

In offline replay, $F(\cdot)$ is not directly observable counterfactually, so we use *regret proxies* that make explicit what is estimated versus what is observed. Concretely, we compute (i) a replay-based value estimate $\widehat{R}(\pi)$ on a discretized action set, and (ii) a model-based value $\sum_t \sum_g \widehat{f}_g(p_t(g), x_t)$ from the fitted response model. We then benchmark a candidate policy against the best policy within the same class found by exhaustive search on the discretized grid using the same estimator; this yields an *in-class* pseudo-regret that is comparable across algorithms even if it is not an oracle regret.

Fairness and governance violations. For any constraint $c_j(p) \le 0$, we track per-period violation and cumulative violation:

$$\mathrm{viol}_{j,t} \;=\; [c_j(p_t)]_+, \qquad \mathrm{Viol}_T \;=\; \sum_{t=1}^{T}\sum_{j=1}^{J}[c_j(p_t)]_+.$$

To separate "rare but large" from "frequent but small" failures, we also report the maximum violation

$$\mathrm{MaxViol}_T \;=\; \max_{t\le T}\max_{j\le J}[c_j(p_t)]_+,$$

and the fraction of periods with any violation. For price-parity governance, we present the realized dispersion

$$\mathrm{Disp}_t \;=\; \max_{g,g'\in\mathcal{G}} |p_t(g) - p_t(g')|,$$

and compare it to Δ. For outcome fairness, we report both the worst-pair disparity $\max_{g,g'} |\mu_g(p_t(g)) - \mu_{g'}(p_t(g'))|$ (in simulation) and its empirical proxy obtained by substituting $\widehat{\mu}^{\log}$ or model-imputed $\widehat{\mu}_g(\cdot)$ (in replay). When a method enforces safety via conservative bounds, we additionally report the *constraint margin* $-U^c_{j,t-1}(p_t)$, which is an operational measure of how close decisions are to the certified boundary.

Volatility, stability, and operational smoothness. To reflect implementation costs, we measure volatility both as a hard-switch count and as total variation:

$$\mathrm{Switch}_T = \#\{t \ge 2 : p_t \ne p_{t-1}\}, \qquad \mathrm{TV}_T = \sum_{t=2}^{T} \|p_t - p_{t-1}\|_1.$$

Beyond feasibility, we summarize *stability* by the distribution of step sizes $\|p_t - p_{t-1}\|_1$ and by run-to-run variability under repeated simulations with different noise seeds (holding contexts fixed). This distinguishes policies that satisfy a budget mechanically from those that produce economically smooth trajectories.

Auditability and statistical reporting. Because governed learning must be explainable ex post, we treat auditability as a measurable output: for each t we log (x_t, p_t), the objective bounds $(L^F_{t-1}(p_t), U^F_{t-1}(p_t))$, the constraint bounds $\{U^c_{j,t-1}(p_t)\}_j$, and (when applicable) dual variables $\{\lambda_{j,t}\}_j$. We then report average posterior widths (a proxy for epistemic uncertainty), empirical coverage of confidence intervals in simulation, and the share of decisions that would remain feasible under a small tightening of thresholds (Δ, ε), which serves as a robustness-to-audit metric. All headline numbers are presented with Monte Carlo standard errors (simulation) or block bootstrap intervals (replay) to reflect time dependence in $\{x_t\}$ and the induced correlation in revenues.

11.3 8. Policy and deployment implications: audit logs from GP posteriors, interpretability, and how to choose fairness/volatility parameters in practice.

A governed pricing system is only as deployable as it is *auditable.* The central operational benefit of a GP-based approach is that every decision comes with a quantitative certificate of what we believed at the time. In deployment we therefore recommend a decision log that is natively aligned with the algorithm: for each period (or block, under batching) we store the chosen vector p_t, the posterior summaries for the objective and each constraint at that point $\big(\mu^F_{t-1}(p_t), \sigma^F_{t-1}(p_t)\big)$ and $\{(\mu^c_{j,t-1}(p_t), \sigma^c_{j,t-1}(p_t))\}_{j=1}^J$, and the derived bounds $\big(U^F_{t-1}(p_t), \{U^c_{j,t-1}(p_t)\}_j\big)$ that justify feasibility and optimality *within the certified action set.* This log should also include the candidate set used by the numerical optimizer (e.g., grid, random restarts), the winning argmax, and a hash of the training data snapshot so that an auditor can reproduce the posterior ex post.

Interpretability then becomes a matter of mapping the argmax rule into statements a regulator or product team can understand. We have found it useful to report a short "reason code" per decision: (i) whether p_t was constrained primarily by parity, by outcome fairness, or by volatility; (ii) the most binding constraint margin $\min_j -U^c_{j,t-1}(p_t)$; and (iii) the opportunity cost of governance, approximated by the gap between the best unconstrained optimistic value and the best safe optimistic value,

$$\mathrm{Cost}_t^{safe} \;\approx\; \max_{p\in[p_\ell,p_h]^G} U^F_{t-1}(p) \;-\; \max_{p\in\mathcal{A}_t^{safe}\cap\mathcal{V}_t} U^F_{t-1}(p).$$

This separates "we did not raise price because it violated fairness" from "we did not change price because the switch budget forbids it," which is essential for internal accountability.

Choosing (Δ, ε) is ultimately a policy question, but we can make the trade-offs legible. In practice we recommend a two-stage procedure. First, feasibility screening: select candidate thresholds that admit a robust interior point (a Slater-type condition) under conservative bounds, i.e., there exists p with $U^c_{j,0}(p) \le -m$ for all j and some margin $m > 0$. This prevents early periods from collapsing to a trivial safe set. Second, value-of-fairness calibration: run an offline stress test (replay or semi-synthetic) to trace a frontier $(\Delta, \varepsilon) \mapsto (\widehat{\mathrm{Rev}_T}, \widehat{\mathrm{Viol}_T})$, and choose the pair at which marginal revenue gains per unit relaxation are no longer compelling. When a primal–dual variant is used, the realized dual variables $\{\lambda_{j,t}\}$ serve as *shadow prices*: persistently large $\lambda_{j,t}$ indicates that the organization is operating at the boundary and should revisit whether the corresponding constraint is appropriately tight or whether measurement noise is being mistaken for disparity.

Volatility parameters (S or V) should be set from operational constraints rather than statistical convenience. A simple translation is to treat each price change as incurring a fixed change-management cost K (engineering, comms, customer support), in which case batching with $B = S + 1$ blocks approximates a solution to a penalized objective $\sum_t F(p_t) - K\,\mathbf{1}\{p_t \neq p_{t-1}\}$. For teams that can tolerate small continuous adjustments but not frequent large jumps, a TV budget V is more faithful; it also yields a monitoring target that can be tracked in real time.

Finally, deployment should incorporate "safety valves" that acknowledge model limitations. Outcome fairness constraints rely on estimable proxies (e.g., conversion), and proxy drift can make a previously certified boundary misleading. We therefore recommend (i) periodic re-estimation of the GP with rolling windows, (ii) drift detectors that trigger a revert-to-p^{safe} mode, and (iii) routine subgroup audits to validate that protected-class definitions and data pipelines match the governance intent. These practices do not eliminate the normative tensions in fairness-aware pricing, but they make the trade-offs explicit, reproducible, and contestable—which is precisely what policy-facing learning systems require.

11.4 9. Conclusion and extensions: contextual pricing, competition, and discrete demand likelihoods.

We studied dynamic, group-dependent pricing under two forms of governance: constraints that encode normative requirements (such as price parity or outcome fairness) and constraints that encode operational limits (such as switch budgets or total variation). Our main message is that these requirements can be integrated *directly* into the exploration–exploitation problem by treating both revenue and constraints as unknown functions over the price space and learning them with a common statistical object—a GP posterior. This yields a transparent decision rule: optimize an optimistic revenue bound over a conservatively feasible (or penalized) action set. The resulting guarantees formalize a pragmatic aspiration in policy-facing learning systems: we can learn what we do not know while maintaining ex ante commitments about what we will not do.

Several extensions are immediate and, in our view, essential for realism. First, *contextual pricing.* In many applications the seller observes time-varying covariates x_t (inventory, macro conditions, user mix) that shift demand. The natural formulation is $f_g(p, x)$ and $c_j(p, x)$, with decisions p_t chosen after observing x_t. One route is a product kernel $k\big((p, x), (p', x')\big) = k_p(p, p')k_x(x, x')$, which preserves GP-UCB-style confidence bounds at the cost of higher information gain Γ_T. Another is a semiparametric model, e.g.,

$$f_g(p, x) = \langle \theta_g, \phi(x) \rangle + h_g(p),$$

combining a low-dimensional context effect with a nonparametric price component. In either case, governance constraints become context-conditional: outcome fairness may be required only within comparable contexts (e.g., within risk bands), which suggests constraints of the form $c_j(p, x) \leq 0$ or constraints averaged over a reference distribution of contexts. The policy implication is subtle: conditioning can reduce spurious disparity findings (by comparing like with like) but can also create loopholes if contexts are themselves correlated with protected status; any contextual extension should therefore be paired with an explicit governance decision about what variables may enter x_t.

Second, *competition.* If multiple sellers set prices simultaneously, the revenue function for a focal seller depends on rivals' prices p_t^{-i}, and governance may interact strategically (e.g., parity constraints can soften competition or, conversely, constrain undercutting). A reduced-form extension treats rivals as part of the context, $x_t = (p_t^{-i}, \text{market signals})$, and asks for regret guarantees relative to a best-response class subject to the same constraints. A more structural extension models a repeated game in which each firm runs a learning algorithm; then the relevant benchmark shifts from regret to equilibrium concepts (coarse correlated equilibrium, Nash in stationary policies) under feasibility constraints. Our safe-set construction remains conceptually useful—it defines actions that are compliant regardless of beliefs about competitors—but formal results will typically require assumptions on how p_t^{-i} evolves (bounded variation, mixing, or oblivious adversaries). From a policy perspective, competition also raises a separate governance layer: constraints that are individually well-intentioned may have market-level effects (price floors, reduced dispersion), so auditing should be complemented with market monitoring.

Third, *discrete demand likelihoods and non-Gaussian feedback.* The baseline analysis uses sub-Gaussian noise on realized revenues, but many platforms observe counts: purchases, clicks, or conversions, often well-modeled by Binomial or Poisson likelihoods. Two practical approaches preserve much of the framework. One is to place a GP prior on a latent utility and use a generalized likelihood (GP classification or log-GP intensity), yielding posterior approximations (Laplace, EP, variational) and replacing closed-form UCBs with calibrated credible bounds. The other is to model $\mu_g(p)$ via a GLM with a nonparametric link in p, which can deliver finite-sample concentration under boundedness conditions. In both cases, constraint learning is often the harder part: fairness metrics are ratios or differences of rates, and careful propagation of uncertainty is required to avoid either unjustified violations or excessive conservatism.

We close with two limitations that are not merely technical. First, governance constraints are only as meaningful as the measurement system that instantiates them; proxy misspecification and drift can turn a formally safe policy into a substantively unfair one. Second, the choice of (Δ, ε) (and the choice of *which* fairness metric to constrain) is normative and cannot be delegated to the algorithm. What our model offers is a disciplined way to surface the opportunity costs and the shadow prices of those normative choices, and a modular platform for extending governed learning to richer settings where context, strategic interaction, and discrete outcomes are unavoidable.