Constrained Dynamic Pricing with Delayed Returns and Carbon Costs: A Low-Dimensional Pareto Frontier and Implementable Pricing Rules

Liz Lemma   Future Detective

January 16, 2026

Abstract

Dynamic pricing work in the source material optimizes a single-objective profit reward using Q-learning under a static elasticity-based demand model. In modern 2026 retail, pricing decisions also shape operational and regulatory outcomes: return flows strain reverse logistics, shipping congestion raises fulfillment costs, and carbon pricing (explicit or contractual) makes emissions a priced input. We propose a clean constrained Markov decision process (CMDP) that augments the pricing state with a one-dimensional delayed-return pipeline and includes platform constraints (price ladders, floors/ceilings) and stability constraints (bounded price moves). Our main theoretical contribution is a frontier result: under mild convexity/concavity conditions, the set of Pareto-efficient tradeoffs among discounted profit, long-run return volume, and long-run emissions is low-dimensional and can be parameterized by two Lagrange multipliers. In a tractable linear-demand/affine-return specification, the optimal policy reduces to a projected affine (threshold-like) pricing rule, providing interpretable prescriptions: constraints enter as shadow prices that raise effective marginal costs of shipment and returns. Empirically, we outline estimation of return propensity as a function of price and context and demonstrate (numerically) that constrained policies can outperform profit-only baselines out-of-sample by reducing return and carbon costs while preserving most margin.

Table of Contents

delay distributions), and carbon pricing mechanisms (fees, internal transfer prices).

3. 3. Model primitives: demand, delayed returns via a pipeline state, congestion costs, carbon charges, and platform/stability constraints; define objectives (profit vs. constrained multi-objective).

4. 4. Markov reduction for delayed returns: show geometric-delay returns imply a one-dimensional sufficient statistic; discuss when higher-dimensional pipelines are needed (flag numerical methods).

5. 5. CMDP formulation and Pareto frontier: define feasible policy set; show convexity of achievable performance vectors; Lagrangian representation and low-dimensional parameterization of the efficient set.

6. 6. Closed-form/tractable pricing rules in a linear model: derive the projected affine optimal pricing rule under linear demand and affine return propensity; interpret multipliers as shadow costs of returns and emissions; provide comparative statics.

7. 7. Learning/estimation and implementation: estimate D and ρ (and emissions factors) from logs; propose a practical algorithm (outer loop over multipliers + inner loop dynamic pricing solver); discuss constraint auditing.

8. 8. Numerical illustration (calibration): simulate or semi-synthetically evaluate against profit-only RL and static OR; show out-of-sample robustness to return-rate shifts and carbon price changes.

9. 9. Extensions: multi-SKU coupling via shared congestion, multi-agent marketplace competition, and heterogeneous customer segments (flag where numerics are required).

10. 10. Policy and platform implications: what constraints are most effective; how carbon taxes/fees shift optimal prices; welfare discussion and auditing recommendations.

11. 11. Conclusion.

1 Introduction and motivation

In 2026, pricing is no longer a single-instrument lever for extracting short-run margin from a demand curve; it is an operational control that reshapes downstream flows in fulfillment networks, reverse logistics, and emissions accounting. Retailers that treat price as "just revenue" often discover that the true unit economics are determined after the sale: by the fraction of orders that return, by when those returns arrive relative to staffing and carrier capacity, and by the shadow cost of carbon embedded in outbound and reverse shipments. We therefore begin from a simple premise: a modern pricing policy must internalize *(i)* delayed returns, *(ii)* congestion and handling costs that are convex in volume, and *(iii)* carbon charges or internal carbon transfer prices that scale with shipping activity. The model we develop is intentionally minimalist—price is the primary action—but it is designed to illuminate the tradeoffs practitioners now face when optimizing across profit, return load, and emissions exposure.

The business motivation is straightforward. First, returns have grown from a nuisance parameter into a strategic determinant of profitability. A high posted price may reduce demand, but it can also screen into (or out of) consumers and use cases with different return propensities; conversely, aggressive promotions can pull forward volume that later reappears as reverse-logistics congestion. Second, fulfillment is increasingly capacity-constrained and surcharge-driven. When warehouses and carriers operate near peak utilization, the marginal cost of shipping an additional unit is not constant: it rises due to labor overtime, cut-off misses, dimensional weight penalties, and carrier peak fees. That nonlinearity is precisely what convex congestion costs are meant to capture, and it converts what would be a static markup problem into an intertemporal control problem: today's price affects not only today's volume, but also tomorrow's operating regime through the return pipeline and through the persistence of capacity stress. Third, carbon accounting has moved from corporate reporting into *priced* constraints. Whether via explicit taxes, sectoral fees, cap-and-trade pass-through, or internal shadow pricing used for capital allocation, emissions now enter the objective function in a way that is operationally similar to a per-unit cost—except that the relevant "tax" can itself be policy-driven and binding at the firm level.

These realities sit uneasily with the canonical dynamic pricing framework taught in operations and economics. In the baseline textbook model, the retailer chooses a price to balance marginal revenue against a constant marginal cost, perhaps under inventory dynamics or demand learning. Even when such models are extended to a dynamic setting with covariates and state dependence, the state is often a *forward* operational object (inventory, capacity, or belief about demand), while the *backward* object—the return pipeline—is either ignored or treated as contemporaneous shrinkage. Yet returns are delayed and clustered: they generate a lag between the sale and its true net revenue, and they impose real costs (inspection, restocking, disposal, customer service) whose timing matters for staffing and facility utilization. A pricing policy that does not “see” this lag may look optimal in-sample and still be systematically fragile out-of-sample, particularly around seasonal peaks when return windows overlap.

Our goal is to keep the model close enough to practice that its policy implications are interpretable, while remaining structured enough to yield clear theoretical lessons. We therefore build around three design choices. The first is a parsimonious representation of delayed returns. Rather than tracking the full age distribution of past sales eligible to return, we use a one-dimensional pipeline state that summarizes the expected mass of pending returns. This is not merely a mathematical convenience; it corresponds to a managerial object (“how much return exposure is in the system?”) and is empirically estimable from aggregate return-window data. The second choice is to model the operational cost of volume via a convex congestion term. This captures the idea that marginal fulfillment cost rises with throughput, even absent binding inventory constraints, and it provides a disciplined way to represent carrier surcharges and warehouse crowding without hard-to-calibrate discontinuities. The third choice is to treat carbon as an explicit component of per-period payoffs and constraints, allowing us to study both price-based regulation (a carbon fee) and quantity-based regulation (an emissions cap) within a unified framework.

These modeling choices also speak directly to the current methodological trend of using reinforcement learning (RL) for pricing. Many deployed systems amount to profit-only learning: a Q-learning or actor–critic agent observes a context vector, chooses a price, receives a scalar reward equal to contemporaneous profit, and updates parameters to improve expected discounted reward. This approach has two well-known failure modes in our setting. First, if the reward is measured at shipment time while returns arrive later, the agent faces a delayed and partially observed penalty; naive Q-learning can then overvalue actions that generate immediate revenue but large future return costs. Second, profit-only rewards do not encode carbon or operational constraints, so the learned policy may be infeasible under emissions budgets or return-handling capacity, even if it performs well on the historical objective. In practice, teams patch these issues with ad hoc penalties, post-processing, or separate “guardrail” rules; our aim is to provide an economic structure that clarifies what those penalties *should* be, and when they can be represented by a small number of shadow prices.

A central theme of the paper is that constraints on returns and emissions are not merely compliance add-ons; they are scarcity constraints that induce implicit prices. When return-processing capacity is limited, an additional expected return has an opportunity cost: it displaces other returns, increases cycle time, or forces outsourcing at higher cost. When emissions are capped or priced, each outbound or reverse shipment consumes a scarce resource. The correct way to incorporate these scarcities is through Lagrange multipliers (shadow costs) that enter the pricing decision exactly like additive marginal costs. This yields a constructive interpretation: we can view "sustainability" and "reverse-logistics" requirements as converting a single-objective pricing problem into a multi-objective problem whose efficient frontier can be explored by tuning a low-dimensional vector of multipliers. For practitioners, this is useful because it suggests that a wide range of corporate policies (a carbon internal price, a return-capacity budget, a service-level target) can be translated into a small set of numbers that a pricing system can ingest and act upon consistently.

We also emphasize implementability. Real platforms rarely allow continuous, unconstrained prices. Retailers face discrete price ladders, floors and ceilings, parity rules across channels, and stability constraints that limit how quickly posted prices may change. These frictions matter for both economics and learning: they induce kinks and projections that can defeat smooth first-order conditions, and they can create persistence in prices that is mistaken for consumer reference effects. Rather than abstracting them away, we incorporate platform feasibility and price-stability constraints as part of the action set, so that the optimal policy is explicitly a *projected* rule: choose the economically preferred price, then map it to the nearest feasible price consistent with platform and stability requirements. This perspective helps bridge theory and deployment, because it separates the core economic logic (the unconstrained "target" price) from the institutional mapping (the feasible posted price).

The resulting framework complements, rather than replaces, standard dynamic pricing and RL formulations. It retains the flexibility of context-dependent demand, allows for stochastic dynamics in demand covariates, and is compatible with model-based estimation or model-free learning. At the same time, it forces us to confront the intertemporal nature of returns and the policy nature of carbon. The contribution is therefore not a new forecasting trick, but a disciplined *control* view: price influences a forward sales flow and a delayed reverse flow, both of which carry congestion and emissions externalities that can be internalized through shadow costs. We will be candid about what this does *not* accomplish. We do not claim that a single scalar pipeline state captures all nuances of heterogeneous return windows, item conditions, or fraud dynamics; nor do we claim that carbon is perfectly proportional to units shipped. Our point is that even a stylized representation is enough to change the qualitative prescription for optimal pricing and to clarify how constraint-aware learning and planning should be structured.

Finally, we highlight the practical question that motivates the rest of the paper: if a retailer must meet a return-volume budget and an emissions budget while operating under platform pricing rules, what does an optimal pricing policy look like, and how can we compute it in a way that is transparent and tunable? The sections that follow introduce the institutional constraints that make the problem nontrivial, formalize the dynamic program with a delayed-return state, and show how a low-dimensional multiplier parameterization recovers the Pareto-efficient tradeoff between profit, returns, and emissions.

2 Institutional background and constraints

Before formalizing primitives, it is useful to be explicit about the institutional frictions that make "optimal pricing" in retail look less like a frictionless markup rule and more like a constrained control problem. In many categories, retailers do not choose a real-valued price each period; they choose a *posted* price within a menu shaped by platform rules, consumer-protection policies, and operational guardrails. Those rules are often motivated by trust and fairness concerns, but they have immediate economic implications: they create nonconvexities (discrete ladders), inertia (stability constraints), and sometimes cross-product coupling (parity and promotional mechanics). Our modeling choice to embed an admissible set and a hard stability bound is therefore not a technical add-on; it mirrors the environment in which algorithmic pricing systems are deployed.

Platform admissibility: floors, ceilings, parity, and ladders. Most large marketplaces and retail channels impose some combination of (i) *price floors* (e.g., MAP agreements, "no-loss" rules tied to wholesale cost, or minimum advertised price policies), (ii) *price ceilings* (e.g., anti–price-gouging restrictions during emergencies, "fair pricing" policies, or category-level caps to prevent extreme outliers), and (iii) *parity constraints* across channels (e.g., the requirement that an on-platform price not exceed the seller's own DTC price, or the platform's use of automatic delisting/suppression when an item is cheaper elsewhere). In addition, posted prices are typically restricted to a *discrete ladder*: currency rounding (e.g., cents), psychological endings (e.g., \$19.99), or platform-defined increments that vary by price range. When promotions are present, the relevant "price" can also be an effective price net of coupons, subscription discounts, or shipping credits; the platform may restrict which combinations are allowed, creating an additional layer of feasibility constraints even if the nominal list price is continuous.

For our purposes, these considerations motivate representing the retailer's action as a choice p_t from an admissible set $\mathcal{P} \subset \mathbb{R}_+$, interpreted broadly to include floors, ceilings, and ladder points. This representation is deliberately agnostic about why a given point is admissible; what matters for the economics is that the retailer cannot freely implement the unconstrained optimizer. In particular, discreteness implies that first-order conditions are at best heuristic: the economically preferred price must be mapped to the nearest feasible posted price, and small changes in context can induce kinks in the policy when the nearest ladder point switches.

Stability constraints and "price change friction" as policy. A second pervasive constraint is *price stability.* Platforms often limit how frequently a seller may change the posted price, or how large a period-to-period change may be, for reasons ranging from consumer trust (avoiding "yo-yo" pricing) to compliance with reference-price rules (preventing artificial list prices that enable perpetual markdown claims). Retailers themselves also impose stability to reduce operational complexity: frequent price changes interact with ad spend, merchandising, call-center scripts, and price-matching commitments. Even when no explicit cap exists, there is frequently an implicit one: large swings can trigger customer-service incidents, social-media backlash, or platform monitoring.

Operationally, stability is often implemented as a hard guardrail: "do not move more than $X\%$ per day" or "do not change price more than once every k days." Our baseline captures the first form as a hard bound

$$|p_t - p_{t-1}| \leq \Delta,$$

intersected with $\mathcal{P}$. This makes the pricing problem inherently dynamic even absent inventories or learning: today's action becomes tomorrow's feasible set. The economic implication is subtle but important. Stability constraints create *state dependence through the previous price* p_{t-1}, which can be mistakenly attributed to consumer reference effects if one does not model the institutional restriction. They also create asymmetry during shocks: when demand covariates move quickly (e.g., sudden traffic spikes), the price cannot jump immediately to the new unconstrained target, so the retailer experiences transient periods of "mispricing" that spill into fulfillment load and, in our setting, into the return pipeline and emissions.

Operational reality of returns: delays, clustering, and capacity. Returns are not contemporaneous with sales. In many categories, the delay distribution is governed by a return window (e.g., 30 days from delivery), shipment lead times, consumer usage/try-on behavior, and carrier pickup or drop-off patterns. As a result, return arrivals are typically *clustered*: peak-season sales can generate an extended reverse-logistics wave weeks later, and that wave interacts with warehouse labor and carrier capacity in ways that are not well approximated by a constant per-unit cost applied at shipment time. Moreover, return rates are not purely exogenous. They vary with product attributes, fit/size uncertainty, marketing channel, and (crucially for pricing) with the type of customer and intended use that a given price point attracts. In practice, teams often observe that deep discounts can increase return propensity (impulse purchases, "bracketing" behavior), while higher prices can either reduce returns (more deliberation) or increase them (higher expectations and dissatisfaction), depending on category.

These realities motivate two modeling choices. First, we treat return propensity as context- and price-dependent, $\rho(p, x) \in [0, 1]$, so that price influences not only demand but also the expected fraction of sold units that will eventually return. Second, we represent delayed arrivals through a pipeline state rather than as immediate shrinkage. This aligns with managerial practice: fulfillment teams monitor outstanding “return exposure” (units that could still come back) and staff accordingly. It also aligns with the needs of algorithmic pricing systems: a pricing policy that ignores pipeline exposure will tend to overvalue actions that generate immediate revenue and postpone costs.

Why a geometric delay is a useful approximation (and when it is not). Real return delays are not literally geometric; hazard rates often change over time (e.g., consumers return shortly after delivery if at all, with a spike near the end of the window). We nonetheless emphasize the geometric specification because it yields a tractable sufficient statistic: a one-dimensional mass of pending returns. Interpreting $\delta \in (0, 1]$ as an *arrival hazard* is often reasonable as an approximation when we aggregate across heterogeneous consumers, shipping speeds, and return channels: mixtures of many idiosyncratic delays can produce an effectively memoryless aggregate in coarse time buckets (e.g., weekly). The approximation is also pragmatic. In deployed systems, the goal is often not to perfectly forecast the age profile of returns, but to incorporate a stable measure of expected near-term return arrivals into pricing and capacity planning.

We will be explicit about the limitation: if the hazard is strongly duration-dependent (e.g., almost no returns until day 25, then a cliff), a scalar pipeline can be insufficient for accurate control, and richer state representations (age bins) become necessary. We return to this point when discussing extensions and the conditions under which closed-form pricing rules cease to be reliable.

Carbon pricing mechanisms in commerce logistics. Carbon enters retail decision-making through a mix of external regulation and internal governance. On the external side, firms may face explicit carbon taxes, cap-and-trade pass-through embedded in carrier rates, or sector-specific fees and fuel surcharges. On the internal side, many firms adopt an internal carbon price used for budgeting, vendor selection, and performance measurement; importantly, this internal price can be binding even when external regulation is not, because it is tied to corporate emissions targets. In both cases, logistics is a natural locus for carbon accounting: outbound shipments and return shipments are measurable events that can be mapped (imperfectly but consistently) to emissions factors, often differentiated by mode (ground vs. air), packaging, distance, and consolidation.

For the purpose of pricing control, what matters is that carbon mechanisms behave like either (i) an additive per-unit charge proportional to emissions, τE_t, or (ii) a quantity constraint (an emissions budget) enforced over a horizon. The former is an explicit "tax" in the objective; the latter is a scarcity constraint that induces a shadow price. In practice, both coexist: a firm may face an internal transfer price τ *and* a corporate cap $\bar{E}$ that triggers escalation when exceeded. Our formulation accommodates this by including direct carbon charges in per-period payoff while also allowing a long-run average emissions constraint.

From institutional detail to model-ready constraints. The common thread across these institutional features is that they convert pricing into a constrained dynamic problem with delayed consequences. Platform feasibility and stability constraints restrict the action set each period and create dependence on the previous posted price. Return delays transform what would be a static "net revenue" adjustment into an intertemporal pipeline that affects future costs, congestion, and (through reverse shipments) emissions. Carbon pricing and carbon budgets convert shipping activity into an explicitly priced (or capped) resource. Taken together, these constraints motivate the state variables and feasibility sets we introduce next, and they clarify why Lagrange multipliers are not merely mathematical artifacts: in practice, they correspond to interpretable shadow costs arising from return-handling capacity and emissions budgets. In the next section, we translate the above institutional elements into primitives—$D(p, x)$, $\rho(p, x)$, a pipeline state s, congestion costs $G(\cdot)$, emissions accounting, and the admissible set $\mathcal{P}$ with stability bound Δ—and we state the retailer's objective as discounted profit subject to long-run constraints on returns and emissions.

3 Model primitives and objective

We now translate the institutional constraints described above into a set of primitives that can be used directly in a dynamic control problem. The goal is not to encode every operational detail of retail logistics, but to isolate the economic channels that matter for pricing: (i) a demand response to posted prices and observable covariates; (ii) delayed, price-dependent returns; (iii) congestion and processing costs that are increasing in shipped and returned volume; (iv) carbon accounting that attaches (possibly regulated) charges or budgets to logistics flows; and (v) feasibility constraints on the posted price reflecting platform admissibility and price stability. Throughout we work in discrete time $t = 0, 1, 2, \ldots$, interpret periods at whatever aggregation is operationally relevant (e.g., day or week), and let $\beta \in (0, 1)$ denote the discount factor.

Demand and context. At the start of each period the retailer observes a vector of demand and operations covariates $x_t \in \mathcal{X}$. This state collects factors such as seasonality, traffic, ad exposure, product attributes, shipping conditions, and any other exogenous shifters of demand or fulfillment performance that are observable at the time the price is chosen. Conditional on (p_t, x_t), expected sales (or outbound shipments) in period t are

$$q_t \;=\; D(p_t, x_t),$$

where $D : \mathbb{R}_+ \times \mathcal{X} \to \mathbb{R}_+$ is assumed to be decreasing in p and sufficiently smooth to support the comparative statics we derive later. We will often impose concavity of $p \mapsto D(p, x)$ or, more directly, concavity of the induced one-period profit in price, but at the primitive level we only require D to be measurable and well-behaved (e.g., bounded on the admissible price range). To keep attention on intertemporal return and emissions effects, we treat x_t as exogenous and evolving according to a Markov kernel $x_{t+1} \sim P(\cdot \mid x_t)$, which captures persistent seasonality and business-cycle variation without introducing strategic interactions.

Returns: propensity and delayed arrival. Each sold unit may eventually return, and returns arrive with a delay. We represent the expected fraction of newly sold units that will eventually return by a return propensity function

$$\rho(p, x) \in [0, 1],$$

allowing price to influence not only the level of demand but also the composition of customers and therefore the expected return rate. Separating ρ from D is useful empirically (the forces that move conversion and the forces that move returns often differ) and economically (pricing can trade off gross margin against downstream reverse-logistics load).

To model delays parsimoniously we assume that, conditional on returning, the return time is geometric with arrival hazard $\delta \in (0, 1]$. Rather than tracking an age distribution of past sales, we summarize the "return exposure" by a single pipeline state $s_t \geq 0$, interpreted as the expected mass of past sales that are still within the return window and have not yet arrived as returns. After choosing p_t and realizing expected sales $q_t = D(p_t, x_t)$, the pipeline updates according to

$$s_{t+1} \;=\; (1 - \delta)s_t + \rho(p_t, x_t)D(p_t, x_t),$$

and expected return arrivals in period t are

$$y_t \;=\; \delta s_t.$$

This formulation makes the economic timing explicit: new sales increase future return exposure, while outstanding exposure decays as returns arrive.

The geometric specification is chosen for tractability and transparency; the next section shows formally why it yields a one-dimensional sufficient statistic and discusses when richer pipeline representations become necessary.

Costs: production, fulfillment, congestion, and return processing. Let c denote the per-unit production or wholesale cost incurred on shipped units, and let f denote the per-unit outbound fulfillment cost (picking, packing, baseline shipping). Returns generate additional operational costs, summarized by a per-unit return processing cost h applied to return arrivals y_t. These terms reflect the accounting reality that many expenses scale approximately linearly with volume, at least locally.

To capture nonlinear capacity pressure we add a congestion cost $G(q)$, where $G : \mathbb{R}_+ \to \mathbb{R}_+$ is convex and increasing. This term can represent warehouse overtime, carrier surcharges, or service-level penalties that escalate when outbound volume is high. Convexity is economically natural (marginal congestion costs rise with load) and technically convenient (it supports concavity of the one-period objective in price under standard demand shapes). While we index congestion by outbound shipments q_t for simplicity, the same approach can be extended to congestion in reverse logistics or to coupled congestion across products; doing so will matter for the numerical extensions we discuss later.

Carbon accounting and carbon charges. Outbound and return shipments generate emissions. We translate these physical flows into a period-t emissions measure

$$E_t \;=\; \kappa_S q_t + \kappa_R y_t,$$

where κ_S and κ_R are emissions factors (e.g., kg CO_2e per outbound unit and per returned unit). These factors may embed average distances, packaging, and mode choice; the point of this reduced form is to make the carbon implication of demand and returns operationally interpretable and directly connected to pricing decisions.

We allow carbon to enter the objective through a per-unit emissions price $\tau \geq 0$, so that carbon charges in period t equal τE_t. This can be interpreted as an external tax, a carrier pass-through, or an internal transfer price used for planning. Importantly, we will also consider a separate long-run emissions cap; in practice a firm may face both a per-unit charge and a budget that becomes binding over longer horizons.

Feasibility: platform admissibility and stability. The retailer does not choose an arbitrary real-valued price. Let $\mathcal{P} \subset \mathbb{R}_+$ denote the admissible set induced by platform rules (floors, ceilings, parity constraints, and ladders). In addition, posted prices are subject to a hard stability constraint:

$$|p_t - p_{t-1}| \leq \Delta,$$

where $\Delta \geq 0$ is a maximum step size. We treat the previous period price p_{t-1} as a state variable p_{-} because it determines the current feasible set. Thus the set of feasible actions in state (x_t, s_t, p_{t-1}) is

$$p_t \in \mathcal{P} \cap [p_{t-1} - \Delta,\ p_{t-1} + \Delta].$$

This intersection captures both discrete admissibility (through $\mathcal{P}$) and dynamic inertia (through the stability interval). The latter is what makes pricing dynamic even absent inventories: today's action restricts tomorrow's feasible set, which in turn shapes the path of sales, returns, and emissions.

One-period profit. We model refunds in a reduced form by treating revenue as accruing only on net non-returned units in the period of return arrival. Given the expected return arrivals $y_t = \delta s_t$, period-t expected net profit from choosing p_t in state (x_t, s_t) is

$$\begin{gathered} \pi_t(p_t; x_t, s_t) = \ p_t\,(q_t - y_t) \ - \ c\,q_t \ - \ f\,q_t \ - \ h\,y_t \ - \ \tau E_t \ - \ G(q_t), \qquad (1) \\ q_t = \ D(p_t, x_t), \qquad y_t = \delta s_t, \qquad E_t = \kappa_S q_t + \kappa_R y_t. \end{gathered}$$

This expression makes the intertemporal channel explicit. Current price affects contemporaneous shipments q_t, which affects contemporaneous congestion and outbound emissions, and it affects future outcomes by pushing new mass into the return pipeline s_{t+1}. Meanwhile, the inherited pipeline s_t creates a current "headwind" through return arrivals y_t, which reduce net revenue and generate processing and carbon costs. Alternative accounting choices—for example, booking refunds at sale rather than at return arrival—can be incorporated by re-indexing the timing of cash flows without changing the underlying state dynamics.

Objectives: profit maximization with long-run constraints. Let μ denote a stationary Markov pricing policy mapping the observable state to a feasible price, $p_t = \mu(x_t, s_t, p_{t-1})$. Our baseline objective is to maximize discounted expected profit subject to long-run average constraints on returns and emissions:

$$\max_{\mu} \quad \mathbb{E}_{\mu}\left[\sum_{t\geq 0} \beta^t \, \pi_t\right] \qquad (2)$$

$$\text{s.t.} \quad \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_{\mu}\left[\sum_{t=0}^{T-1} y_t\right] \leq \bar{R}, \qquad \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_{\mu}\left[\sum_{t=0}^{T-1} E_t\right] \leq \bar{E}, \qquad (3)$$

$$p_t \in \mathcal{P} \cap [p_{t-1} - \Delta,\ p_{t-1} + \Delta] \quad \forall t.$$

We interpret $\bar{R}$ as a return-handling capacity (or a service-level guardrail on reverse logistics) and $\bar{E}$ as an emissions budget (external or internal).

Equivalently, one may view the problem as a multi-objective trade-off among (expected) profit, returns, and emissions; the constrained formulation emphasizes that in many organizations the latter two are set as targets or budgets rather than as objectives chosen endogenously by a profit center. In the sections that follow we will rely on a Lagrangian relaxation of (3) to describe the Pareto frontier with a small number of shadow prices, but the primitives above are the foundation: a demand system, a delayed-return pipeline, convex operating costs, carbon accounting, and platform/stability feasibility.

4 Markov reduction for delayed returns

Delayed returns create an immediate modeling tension. On the one hand, the economics are inherently intertemporal: today's price influences not only today's shipments but also the timing and volume of future return arrivals. On the other hand, standard dynamic programming techniques rely on a low-dimensional state that is *Markov*, i.e., summarizes all payoff-relevant information from the past. If return arrivals depended on the entire history of sales with rich duration dependence, then the retailer would, in principle, need to carry forward an ever-growing record of past cohorts in order to forecast future reverse-logistics load. The purpose of our geometric-delay assumption is to show that this complexity is not inevitable: with constant return hazard, delayed returns admit a one-dimensional sufficient statistic.

From cohort dynamics to a pipeline state. To see the issue transparently, imagine indexing sales cohorts by their sale date. Let

$$r_t \;\equiv\; \rho(p_t, x_t)\, D(p_t, x_t)$$

denote the expected number of units sold at time t that will *eventually* return (not necessarily immediately). Under geometric delays with hazard δ, each unit in this "eventual-return" cohort returns in the next period with probability δ, and otherwise remains pending with probability $1 - \delta$. Hence, conditional on r_t, the expected return arrivals in period $t + k$ generated by cohort t are

$$\mathbb{E}[\,\text{arrivals at } t+k \text{ from cohort } t \mid r_t\,] \;=\; \delta(1-\delta)^{k-1}\, r_t, \qquad k = 1, 2, \ldots$$

This expression highlights the source of tractability: the entire path of expected future arrivals from a cohort is a geometric tail with a single parameter δ.

Define the (expected) return pipeline at time t as the discounted sum of past eventual-return cohorts that have not yet arrived:

$$s_t \;\equiv\; \sum_{j=0}^{\infty}(1-\delta)^j\, r_{t-1-j} \;=\; \sum_{j=0}^{\infty}(1-\delta)^j\, \rho(p_{t-1-j}, x_{t-1-j})D(p_{t-1-j}, x_{t-1-j}). \tag{4}$$

Interpreted literally, s_t is the expected mass of units sold in the past that are still "at risk" of returning as of period t: recent cohorts enter the pipeline with weight 1, older cohorts persist with survival weight $(1-\delta)^j$.

Proposition 1 (one-dimensional sufficient statistic). Under the geometric-delay assumption, the scalar s_t is sufficient to forecast expected return arrivals and to update future return exposure. In particular,

$$y_t \;=\; \delta s_t, \tag{5}$$

and

$$s_{t+1} \;=\; (1-\delta)s_t \;+\; \rho(p_t, x_t)D(p_t, x_t). \tag{6}$$

Proof sketch. Start from (4). Return arrivals at time t are the hazard δ applied to the mass currently pending, which yields (5). For the recursion, observe that between t and $t+1$, the pipeline shrinks by the survival factor $(1-\delta)$ and then receives the new eventual-return cohort $r_t = \rho(p_t, x_t)D(p_t, x_t)$. Formally,

$$s_{t+1} = \sum_{j=0}^{\infty}(1-\delta)^j r_{t-j} = r_t + (1-\delta)\sum_{j=0}^{\infty}(1-\delta)^j r_{t-1-j} = r_t + (1-\delta)s_t,$$

which is exactly (6). □

Markov property and economic interpretation. The recursion (6) is the key to the Markov reduction. It implies that, for purposes of expected profits and constraints, the retailer need not track the entire sales history: all payoff-relevant implications of past pricing for future return arrivals are summarized by the single number s_t. Combined with the Markov evolution of covariates $x_{t+1} \sim P(\cdot \mid x_t)$ and the stability constraint that makes p_{t-1} payoff-relevant via feasibility, the augmented state

$$(x_t,\; s_t,\; p_{t-1})$$

is Markov. Economically, this state has a clean operational meaning: x_t captures contemporaneous demand and fulfillment conditions, p_{t-1} captures platform-induced inertia, and s_t captures the inherited "return headwind" that will translate into near-term refund/processing/emissions burden through $y_t = \delta s_t$.

The memorylessness embedded in the geometric hazard is doing all of the work. It says that, among units still pending return, the probability of arriving next period is independent of their age. This can be viewed as a reduced-form approximation to settings where returns are processed with a roughly constant weekly rate, or where shipment/processing delays are the dominant source of variation rather than consumer deadline behavior. It is also the natural discrete-time analog of exponential waiting times.

When one dimension is not enough. The one-dimensional pipeline is not a universal truth; it is a disciplined consequence of the constant-hazard assumption. In many retail categories, return behavior exhibits strong duration dependence: return rates may spike near the end of a posted return window, or may be front-loaded due to rapid try-on and immediate dissatisfaction. More generally, suppose the conditional probability that a pending unit returns next period depends on its age k (periods since sale), with hazards δ_k that are not constant. Then two histories that generate the same scalar $\sum_j w_j r_{t-1-j}$ can nevertheless imply different future arrival paths because the age composition differs. In such cases, s_t is no longer sufficient: the retailer must track an *age distribution* of pending returns.

A convenient way to represent this is to maintain an age-binned vector state. For instance, if returns can occur only within a finite window of length L, one can define $s_t^{(k)}$ as the expected mass of eventual-return units that are currently age $k \in \{1, \dots, L\}$ and have not yet arrived. The pipeline update becomes a deterministic "shift" plus inflow from new sales, and expected arrivals are a weighted sum of the bins:

$$y_t \;=\; \sum_{k=1}^{L} \delta_k \, s_t^{(k)}, \qquad s_{t+1}^{(1)} = r_t, \qquad s_{t+1}^{(k+1)} = (1-\delta_k) s_t^{(k)} \;\; (k = 1, \dots, L-1).$$

The state dimension is then L rather than 1. This is still Markov, but it is computationally more demanding, and it blurs the transparent interpretation of a single return-exposure index.

A related failure of the scalar reduction occurs if the return propensity itself depends on cohort attributes that are not adequately summarized by (x_t, p_t) at the time of sale. For example, if promotions attract systematically different consumers whose return timing differs (not just their eventual-return probability), then the pipeline needs to distinguish cohorts by those latent types. In practice one may address this by enriching the observed context x_t, but when heterogeneity is unobserved the pipeline state must absorb it, again raising dimensionality.

Implications for analysis and the role of numerics. Once the pipeline becomes high-dimensional, closed-form pricing rules are typically unavailable. The fundamental control problem remains a (constrained) Markov decision process, but solving it exactly requires dynamic programming over a larger state space, and the stability constraint $|p_t - p_{t-1}| \leq \Delta$ further couples decisions over time.

This is the point where numerical methods become essential. Two pragmatic approaches are especially useful in the present setting. First, one can *approximate* non-geometric delays by a small mixture of geometric components (a discrete-time analog of phase-type approximations). Concretely, if the return-time distribution can be well-approximated by a mixture $\sum_{m=1}^{M} \omega_m \mathrm{Geom}(\delta_m)$, then the pipeline can be represented by an M-dimensional vector $s_t = (s_t^{(1)}, \ldots, s_t^{(M)})$ with the same linear recursion as (6) applied componentwise. This retains much of the interpretability of the scalar pipeline while capturing richer duration patterns. Second, one can directly solve the resulting higher-dimensional problem using approximate dynamic programming or simulation-based methods (e.g., value-function approximation over $(x_t, s_t^{(1)}, \ldots, s_t^{(L)}, p_{t-1})$), which is often feasible in modern retail data environments where transitions can be simulated and policies can be evaluated offline.

Our broader message is therefore twofold. The geometric model is not merely a mathematical convenience: it isolates a case where delayed returns admit a parsimonious state, allowing us to transparently connect pricing incentives to downstream reverse-logistics and carbon costs. At the same time, when the institutional reality requires richer timing (finite windows, deadline effects, cohort heterogeneity), the same economic structure carries through, but the state must expand and numerical tools become the natural complement to theory. The next step is to express the retailer's problem as a constrained MDP and to show how the trade-off among profit, returns, and emissions can still be organized by a low-dimensional set of shadow prices even when the underlying dynamics are complex.

5 CMDP formulation and a low-dimensional Pareto frontier

Having established that delayed returns admit a Markov representation in the augmented state $z_t \equiv (x_t, s_t, p_{t-1})$, we can now state the retailer's problem in the language of a *constrained Markov decision process* (CMDP). This step is conceptually useful for two reasons. First, it cleanly separates *what is controllable* (the pricing rule) from *what is regulated or capacity-limited* (average returns and emissions). Second, it reveals that the trade-off among profit, reverse-logistics load, and carbon impact can be organized by a small number of *shadow prices*—a point that will directly motivate implementable pricing heuristics in the linear specification that follows.

Admissible policies and feasibility. At each state $z = (x, s, p_-)$, the platform and stability constraints restrict feasible prices to the correspondence

$$\mathcal{A}(z) \;\equiv\; \mathcal{P} \;\cap\; [\,p_- - \Delta,\; p_- + \Delta\,].$$

A (stationary) Markov pricing policy is a measurable mapping

$$\mu :\; (x, s, p_-) \;\mapsto\; \Delta(\mathcal{A}(x, s, p_-)),$$

where $\Delta(\cdot)$ denotes the set of probability distributions over actions. Allowing randomization is technically convenient in the CMDP and, economically, can be interpreted as the retailer mixing among nearby ladder prices when the platform discretization is coarse. The induced controlled Markov chain evolves as

$$x_{t+1} \sim P(\cdot \mid x_t), \qquad s_{t+1} = (1-\delta)s_t + \rho(p_t, x_t)D(p_t, x_t), \qquad p_t \sim \mu(\cdot \mid x_t, s_t, p_{t-1}).$$

Let $\pi(p; x, s)$ denote the one-period expected net profit (including congestion and carbon charges) evaluated at $q = D(p, x)$ and $y = \delta s$. For any stationary policy μ and initial state z_0, define the discounted profit objective

$$J(\mu; z_0) \;\equiv\; \mathbb{E}_\mu\Big[\sum_{t\geq 0} \beta^t\, \pi(p_t; x_t, s_t) \,\Big|\, z_0\Big],$$

and the long-run average resource consumptions

$$R(\mu; z_0) \;\equiv\; \limsup_{T\to\infty} \frac{1}{T}\,\mathbb{E}_\mu\Big[\sum_{t=0}^{T-1} y_t \,\Big|\, z_0\Big], \qquad E(\mu; z_0) \;\equiv\; \limsup_{T\to\infty} \frac{1}{T}\,\mathbb{E}_\mu\Big[\sum_{t=0}^{T-1} E_t \,\Big|\, z_0\Big].$$

We call μ *feasible* if it respects the per-period action constraints and satisfies $R(\mu; z_0) \leq \bar{R}$ and $E(\mu; z_0) \leq \bar{E}$. In many applications one works under standard ergodicity conditions ensuring that $R(\mu; z_0)$ and $E(\mu; z_0)$ do not depend on z_0 (or depend only through transient effects); we do not need to insist on this here, but it clarifies interpretation when the system reaches a steady operating regime.

The achievable performance set is convex. Consider the set of attainable performance vectors

$$\mathcal{V} \;\equiv\; \Big\{\big(J(\mu; z_0),\; R(\mu; z_0),\; E(\mu; z_0)\big) : \mu \text{ stationary Markov}\Big\} \subset \mathbb{R} \times \mathbb{R}^2_+.$$

A key structural fact is that $\mathcal{V}$ is convex once we allow randomized policies. The economic content is simple: if we can implement two stationary operating modes, then we can also implement their probabilistic mixture.

Formally, fix two stationary Markov policies μ^1, μ^2 and let $\theta \in [0, 1]$. Construct a mixed policy μ^θ that draws $I \sim \text{Bernoulli}(\theta)$ at time 0 and then follows μ^1 forever if $I = 1$ and μ^2 forever if $I = 0$. Because expectations are linear in this initial randomization,

$$J(\mu^\theta; z_0) = \theta J(\mu^1; z_0) + (1 - \theta)J(\mu^2; z_0),$$

and likewise $R(\mu^\theta; z_0)$ and $E(\mu^\theta; z_0)$ are the corresponding convex combinations. Hence $\mathcal{V}$ is convex. Under mild boundedness/compactness conditions (e.g., bounded prices, bounded costs, and continuity of primitives), one also obtains closedness/compactness of $\mathcal{V}$, which ensures that efficient frontier problems admit solutions.

This convexity is more than a technicality: it is precisely what justifies interpreting the profit–returns–emissions trade-off as a *frontier* that can be traced out by varying shadow prices. Without convexity, scalarization methods would generally miss efficient points.

Pareto efficiency and scalarization. We say that a vector $(J, R, E) \in \mathcal{V}$ is *Pareto efficient* (with J to be maximized and R, E to be minimized) if there is no other attainable triple $(J', R', E') \in \mathcal{V}$ with $J' \geq J$, $R' \leq R$, $E' \leq E$ and at least one inequality strict. The constrained optimization problem,

$$\max_{\mu} \; J(\mu; z_0) \quad \text{s.t.} \quad R(\mu; z_0) \leq \bar{R}, \quad E(\mu; z_0) \leq \bar{E},$$

selects a particular efficient point corresponding to the caps $(\bar{R}, \bar{E})$. More broadly, the entire efficient set can be recovered via weighted-sum scalarization. Under Slater-type feasibility (existence of a strictly feasible policy with slack in both constraints), standard CMDP duality implies that for any efficient point there exists a pair of nonnegative multipliers $\lambda = (\lambda_R, \lambda_E) \in \mathbb{R}^2_+$ that support it.

Concretely, consider the Lagrangian-relaxed objective

$$L(\mu, \lambda; z_0) \equiv J(\mu; z_0) \; - \; \lambda_R\big(R(\mu; z_0) - \bar{R}\big) \; - \; \lambda_E\big(E(\mu; z_0) - \bar{E}\big). \qquad (7)$$

For any fixed $\lambda \geq 0$, maximizing $L(\mu, \lambda; z_0)$ over policies μ is an *unconstrained* MDP with a modified per-period reward that subtracts penalties for return arrivals and emissions (and adds the constants $\lambda_R\bar{R} + \lambda_E\bar{E}$, which do not affect the optimizing policy). In particular, if we write the per-period λ-penalized payoff as

$$\tilde{\pi}_\lambda(p; x, s) \;\equiv\; \pi(p; x, s) - \lambda_R\, y - \lambda_E\, E, \qquad y = \delta s, \quad E = \kappa_S D(p, x) + \kappa_R \delta s,$$

then the corresponding Bellman equation takes the standard form

$$V_\lambda(x, s, p_-) = \max_{p \in \mathcal{A}(x,s,p_-)} \Big\{ \tilde{\pi}_\lambda(p; x, s) + \beta\, \mathbb{E}\big[V_\lambda(x', s', p) \mid x, s, p\big] \Big\},$$

with $s' = (1 - \delta)s + \rho(p, x)D(p, x)$ and $x' \sim P(\cdot \mid x)$. The crucial dimensionality observation is that λ is only two-dimensional: regardless of the complexity of the state dynamics, the frontier is indexed by at most two scalars because there are only two long-run constraints.

KKT conditions and the multiplier interpretation. When strong duality holds (as it does under the convexity/Slater conditions described above), there exists an optimal constrained policy $\mu^\star$ and multipliers $\lambda^\star \geq 0$ such that $\mu^\star$ solves the unconstrained $\lambda^\star$-penalized MDP, and complementary slackness holds:

$$\lambda_R^\star\big(R(\mu^\star; z_0) - \bar{R}\big) = 0, \qquad \lambda_E^\star\big(E(\mu^\star; z_0) - \bar{E}\big) = 0.$$

Thus, if the return-volume cap is slack, then $\lambda_R^\star = 0$ and the retailer behaves as if returns were unconstrained; if the cap binds, then $\lambda_R^\star > 0$ and the optimal policy internalizes an additional per-unit cost of returns. The same logic applies to emissions. This is the rigorous sense in which multipliers act like *shadow prices* of scarce reverse-logistics and emissions capacity, and it is what allows us to translate regulatory or operational constraints into pricing incentives.

Practical implication: tracing the frontier by tuning two numbers. From an implementation perspective, the multiplier parameterization is attractive: rather than solving a family of constrained dynamic programs directly, we can solve a sequence of unconstrained MDPs indexed by (λ_R, λ_E) and then adjust multipliers until the induced policy meets the desired average caps. In practice, this can be done with simple outer-loop schemes (e.g., subgradient updates on λ) wrapped around an inner-loop dynamic program or simulation-based policy evaluation. The approach is not without limitations—nonconvexities, poor state aggregation, or binding integer ladder constraints can complicate convergence and may create local irregularities—but it provides a disciplined organizing device for both computation and economic interpretation.

In the next section we specialize to a linear demand/affine return-propensity environment. There, the Lagrangian formulation not only certifies that two multipliers suffice to span the efficient set, but also yields tractable projected pricing rules that make the shadow-cost logic operational.

6 6. Closed-form/tractable pricing rules in a linear model: derive the projected affine optimal pricing rule under linear demand and affine return propensity; interpret multipliers as shadow costs of returns and emissions; provide comparative statics.

We now specialize the primitives to a linear-demand, affine-returns environment in which the Bellman maximization admits a transparent first-order condition and, after accounting for platform and stability restrictions, an implementable *projected affine* price rule. The point of this exercise is not that all realistic demand and return mechanisms are linear, but that linearity isolates the economic forces in a way that can be used as (i) a baseline policy, (ii) a local approximation around an operating point, or (iii) an interpretable parametric policy class in richer numerical solutions.

Linear specification and the λ-penalized one-period objective. Fix a context x and write

$$D(p,x) = a(x) - bp, \qquad b > 0,$$

and

$$\rho(p,x) = \rho_0(x) + \rho_1 p,$$

with the understanding that $\rho(\cdot,x)$ is clipped to $[0,1]$ in implementation (we work on the interior region where the clipping does not bind, so derivatives are well defined). Given pipeline state s, return arrivals are $y = \delta s$, and emissions are $E = \kappa_S(a - bp) + \kappa_R \delta s$. For fixed multipliers $\lambda = (\lambda_R, \lambda_E)$, the per-period term entering the λ-relaxed Bellman operator is

$$\tilde{\pi}_\lambda(p;x,s) = p\big(D(p,x) - \delta s\big) - (c+f)D(p,x) - h\,\delta s - (\tau + \lambda_E)\big(\kappa_S D(p,x) + \kappa_R \delta s\big) - \lambda_R\,\delta s - G(D(p$$

up to additive constants $\lambda_R \bar{R} + \lambda_E \bar{E}$ that do not affect choice. Two features are worth emphasizing. First, the emissions multiplier λ_E simply *adds* to the prevailing carbon price τ, so outbound shipments behave as if they faced the augmented carbon price $\tau + \lambda_E$. Second, because our reduced-form profit writes net revenue as $p(q - y)$, a larger outstanding pipeline s increases the contemporaneous refund exposure term $p\,\delta s$, creating a direct channel through which the state s shifts the preferred price even before we account for how p influences future pipeline accumulation.

A closed-form benchmark (ignoring continuation and treating G as locally linear). To obtain a simple baseline, suppose we approximate the continuation value as locally flat in p and s and either ignore congestion or treat $G'(q)$ as approximately constant around the relevant volume range.[1] Let $q = a - bp$. Differentiating $\tilde{\pi}_\lambda$ with respect to p yields

$$\frac{\partial \tilde{\pi}_\lambda}{\partial p} = \underbrace{(q - \delta s) + p\frac{dq}{dp}}_{\text{net revenue margin}} + \underbrace{\big(c + f + (\tau + \lambda_E)\kappa_S\big)\Big(-\frac{dq}{dp}\Big)}_{\text{effective marginal cost}} + \underbrace{G'(q)\Big(-\frac{dq}{dp}\Big)}_{\text{congestion pass-through}} = (a - \delta s) - 2bp + b\tilde{c}_S +$$

[1] This is a standard "local rule" interpretation: when the value function varies slowly relative to current profits, or when we use the rule as a one-step lookahead heuristic, the dominant comparative statics are already visible.

where $\tilde{c}_S \equiv c+f+(\tau+\lambda_E)\kappa_S$ is the effective per-unit outbound marginal cost including carbon charges and the emissions shadow price. Dropping the $G'(q)$ term (or absorbing a constant approximation into $\tilde{c}_S$), the unconstrained maximizer solves

$$\hat{p}_\lambda(x,s) = \frac{a(x)-\delta s}{2b} + \frac{\tilde{c}_S}{2}.$$

This expression is already informative: higher demand intercept $a(x)$ increases the price, greater price sensitivity b compresses markups, a larger return pipeline s pushes the price down through refund exposure, and a higher effective carbon cost $(\tau+\lambda_E)\kappa_S$ passes through at rate $\kappa_S/2$ in this stylized benchmark.

Incorporating congestion and the dynamic pipeline channel. The previous rule abstracts from the fact that today's price also affects tomorrow's pipeline via

$$s' = (1-\delta)s + \rho(p,x)D(p,x).$$

To see how λ_R and λ_E enter beyond the outbound-shipment channel, it is enough to recognize that (i) an incremental increase in s' raises expected future return arrivals, return-processing costs, and return-shipping emissions, and (ii) these future burdens are exactly the objects priced by λ_R and λ_E. A tractable approximation is to posit that, in the relevant region, the λ-value function is approximately affine in the pipeline state,

$$V_\lambda(x,s,p_-) \approx \bar{V}_\lambda(x,p_-) - \eta_\lambda(x,p_-)\,s, \qquad \eta_\lambda(x,p_-) \geq 0,$$

so that η_λ is the (state-dependent) shadow value of reducing the pipeline by one unit. Then the continuation term contributes $\beta V_s \frac{\partial s'}{\partial p} = -\beta\eta_\lambda \frac{\partial s'}{\partial p}$ to the first-order condition. Under the affine ρ and linear demand,

$$\frac{\partial s'}{\partial p} = \frac{\partial}{\partial p}\big[(\rho_0(x)+\rho_1 p)(a(x)-bp)\big] = \rho_1 a(x) - b\rho_0(x) - 2b\rho_1 p.$$

If we additionally take $G(q) = \frac{g}{2}q^2$ (so $G'(q) = gq$), the first-order condition remains linear in p, and we obtain the unconstrained affine rule

$$\hat{p}_\lambda(x,s) = \frac{a(x) - \delta s + b\tilde{c}_S + bga(x) - \beta\eta_\lambda\,\rho_1 a(x) + \beta\eta_\lambda\,b\rho_0(x)}{2b + b^2 g - 2\beta\eta_\lambda\,b\rho_1}.$$

The additional terms have a clean interpretation: when η_λ is large (returns/emissions capacity is scarce in present value), the retailer values lowering $\rho(p,x)D(p,x)$, which is achieved by increasing p; hence λ_R and λ_E raise prices *through* their effect on η_λ even though they do not directly enter the myopic margin on new sales.

Projection onto platform ladders and stability bands. The economically preferred price $\hat{p}_\lambda(x, s)$ is only a target: the platform and stability constraints restrict feasible actions at state $z = (x, s, p_-)$ to $\mathcal{A}(z) = \mathcal{P} \cap [p_- - \Delta,\ p_- + \Delta]$. The implementable stationary rule is therefore the projection

$$p^\star_\lambda(x, s, p_-) = \Pi_{\mathcal{A}(x,s,p_-)}\big(\hat{p}_\lambda(x, s)\big),$$

where Π denotes the nearest feasible point (for a continuous interval this is clipping; for a discrete ladder it is "rounding" to the nearest admissible rung, with ties broken consistently). This formula makes the role of Δ operational: when Δ is small, the projection binds frequently and the policy behaves like a smoothed version of the unconstrained target, adjusting only gradually to shocks in $a(x)$, in the pipeline s, or in shadow costs.

Shadow-cost interpretation of λ_R and λ_E. In the linear rule, λ_E has an immediate outbound channel through $\tilde{c}_S = c + f + (\tau + \lambda_E)\kappa_S$: raising λ_E increases the effective marginal cost of shipments and increases the target price. The return multiplier λ_R operates primarily through the continuation term: by raising the present value cost of adding to the pipeline via $\rho(p, x)D(p, x)$, it increases η_λ and thereby steepens the incentive to price away from high-return volume. Symmetrically, the emissions multiplier also loads onto the pipeline channel because additional future returns carry return-shipping emissions κ_R and thus become more expensive when λ_E is high. In this sense, (λ_R, λ_E) are precisely the numbers that translate long-run caps into per-unit, forward-looking "capacity prices" on reverse logistics and carbon.

Comparative statics. The projected-affine structure yields immediate directional predictions (holding fixed the projection region so that derivatives are taken with respect to the unconstrained target). In the benchmark rule $\hat{p}_\lambda = \frac{a-\delta s}{2b} + \frac{\tilde{c}_S}{2}$, we have

$$\frac{\partial \hat{p}_\lambda}{\partial a} = \frac{1}{2b} > 0, \qquad \frac{\partial \hat{p}_\lambda}{\partial s} = -\frac{\delta}{2b} < 0, \qquad \frac{\partial \hat{p}_\lambda}{\partial \tau} = \frac{\kappa_S}{2} > 0, \qquad \frac{\partial \hat{p}_\lambda}{\partial \lambda_E} = \frac{\kappa_S}{2} > 0,$$

and $\partial \hat{p}_\lambda / \partial b < 0$ in the sense that greater elasticity compresses the intercept and cost pass-through. In the dynamic expression, the same signs persist, with two refinements: (i) larger ρ_1 (returns more sensitive to price) increases the marginal benefit of price as a return-management instrument, amplifying the impact of η_λ and thus of (λ_R, λ_E); and (ii) larger δ strengthens the immediate refund channel (more returns arrive now), making s more influential in current pricing. Finally, the stability bound Δ does not change the target $\hat{p}_\lambda$ but increases the frequency of clipping/rounding; in periods with large shocks to $a(x)$ or large movements in η_λ, tighter Δ mechanically slows adjustment and can force temporary violations of the unconstrained frontier direction (e.g., price remaining “too low” when λ_E rises), which is exactly why the projection representation is useful for auditing and for anticipating when constraints will bind.

Learning primitives from operational logs. To take the projected-affine logic to data, we need estimates of the primitives that govern (i) contemporaneous demand, $D(p, x)$, (ii) the return propensity for newly sold units, $\rho(p, x)$, (iii) the return-delay parameter δ, and (iv) the emissions factors (κ_S, κ_R) (and, if modeled, the congestion function G). The practical advantage of our Markov reduction is that delayed returns can be learned without maintaining a full age distribution: for any candidate (ρ, δ), the pipeline state can be updated online via

$$s_{t+1} = (1 - \delta)s_t + \rho(p_t, x_t)\, D(p_t, x_t),$$

so estimation can be organized around predicting two observables from logs: sales q_t and realized return arrivals (or return initiations) y_t.

For demand, we recommend starting with a model class that is flexible in x but disciplined in p. In many retail settings a generalized linear model (e.g., log-link Poisson or negative binomial for units) with a price term and rich controls captures most variation:

$$\mathbb{E}[q_t \mid p_t, x_t] = D(p_t, x_t) = \exp\{\phi(x_t) - \psi(x_t)\, p_t\},$$

or, if one prefers to adhere to the linear benchmark used for intuition, $D(p, x) = a(x) - bp$ with $a(x)$ estimated by regression or by a supervised learner mapping $x \mapsto a(x)$. The central econometric caveat is price endogeneity: prices are chosen in response to anticipated demand shocks, so naive regression can bias price sensitivity. In practice, we can mitigate this in at least three complementary ways: (i) exploit randomized price experiments (even small within-band perturbations are useful under Δ); (ii) use instrumental variables based on cost, inventory, or platform-driven shocks that shift p_t but are plausibly independent of latent demand; and (iii) fit a demand model jointly with the pricing policy in a policy-evaluation framework (e.g., doubly robust / orthogonalized estimators) to reduce sensitivity to policy-induced selection.

For returns, we separate *propensity* from *timing*. Given sales q_t, we model the expected number of units from period t that will eventually return as $\rho(p_t, x_t)\, q_t$, with ρ estimated from order-level labels (returned vs. not returned) using logistic regression or a calibrated classifier:

$$\Pr(\text{return} \mid p, x) = \rho(p, x) = \sigma\big(\theta^\top \varphi(p, x)\big),$$

where $\varphi(p, x)$ includes price and the return-relevant covariates (size/fit signals, category, customer segment, shipping method). When ρ is allowed to depend on p, we again face endogeneity: low prices may attract marginal customers with different return behavior. The same experimental/IV logic applies, and, operationally, it is often sufficient to instrument only the price component while controlling flexibly for x.

For timing, we estimate the geometric hazard δ from observed return delays (order date to return arrival). If T is the discrete delay in periods, the geometric model implies $\Pr(T = k) = (1 - \delta)^{k-1}\delta$; the MLE is $\hat{\delta} = 1/\bar{T}$ when truncation is negligible. With censoring (units still in-window), we can use a standard survival likelihood. Although geometric timing is stylized, it is often a good operational approximation over weekly horizons; importantly, deviations from geometricity can be detected by goodness-of-fit diagnostics on delay histograms, at which point Proposition 5 tells us what state augmentation is required.

Finally, emissions factors κ_S and κ_R can be estimated from shipment telemetry (carrier, zone, distance, weight) mapped through a transportation emissions model, or taken from life-cycle accounting tables. In implementation we typically treat them as product–lane averages (SKU × origin × destination region), and we carry uncertainty bands because auditing a carbon cap requires a conservative accounting stance. The same infrastructure can produce a realized emissions series $E_t = \kappa_S q_t + \kappa_R y_t$ consistent with our per-period objective.

From estimates to a deployable policy: an outer–inner loop design. With estimated primitives in hand, we implement pricing via a two-level procedure that mirrors the theory: an *inner loop* solves the Lagrangian-relaxed MDP for fixed multipliers $\lambda = (\lambda_R, \lambda_E)$, and an *outer loop* adjusts λ until the long-run average constraints on returns and emissions are met (up to a safety buffer). This design is appealing because it cleanly separates (i) solving a standard discounted-control problem from (ii) enforcing long-run resource constraints, and it preserves interpretability: λ_R and λ_E are the learned shadow prices of reverse-logistics capacity and carbon capacity.

Concretely, for any candidate λ, the inner-loop problem is an unconstrained MDP with state $z = (x, s, p_-)$, feasible action set $\mathcal{A}(z) = \mathcal{P} \cap [p_- - \Delta, p_- + \Delta]$, reward $\tilde{\pi}_\lambda(p; x, s)$, and transition

$$x' \sim P(\cdot \mid x), \qquad s' = (1 - \delta)s + \rho(p, x)D(p, x), \qquad p'_- \equiv p.$$

We can solve this inner loop by value iteration or policy iteration after discretizing s and (if needed) coarsening x to a finite state representation. In settings where the linear specification is a reasonable approximation, we can instead fit a parametric target rule $\hat{p}_\lambda(x, s)$ (affine in learned features of x and in s) and deploy the projected policy $p = \Pi_{\mathcal{A}(z)}(\hat{p}_\lambda(x, s))$; this yields a fast, stable controller that naturally respects platform and stability constraints.

The outer loop then updates λ using a primal–dual or stochastic approximation step based on measured (or simulated) average constraint usage. Let the per-period resource consumptions be $g_R(z,p) = y = \delta s$ and $g_E(z,p) = E = \kappa_S D(p,x) + \kappa_R \delta s$. Given a policy μ_λ from the inner loop, we estimate its long-run averages $\bar{g}_R(\lambda)$ and $\bar{g}_E(\lambda)$ (via stationary simulation under the estimated dynamics, or from online rollouts), and update

$$\lambda_R^{k+1} = \Big[\lambda_R^k + \alpha_k\big(\bar{g}_R(\lambda^k) - \bar{R}\big)\Big]_+, \qquad \lambda_E^{k+1} = \Big[\lambda_E^k + \alpha_k\big(\bar{g}_E(\lambda^k) - \bar{E}\big)\Big]_+,$$

with step sizes $\alpha_k \downarrow 0$. Intuitively, if a candidate policy exceeds the return cap, we raise λ_R so that the next inner-loop solution prices returns more aggressively; similarly for emissions and λ_E. In many applications, this outer loop converges quickly because there are only two multipliers and because the mapping $\lambda \mapsto (\bar{g}_R(\lambda), \bar{g}_E(\lambda))$ is typically monotone in the relevant range.

Operational details: state tracking, exploration, and robustness. A practical deployment requires real-time tracking of s_t. Under our recursion this is straightforward: we maintain s_t deterministically from past prices and predicted sales, optionally correcting it with realized return arrivals. One useful filter is

$$s_{t+1} \leftarrow (1-\delta)s_t + \rho(p_t, x_t)\,\hat{q}_t \quad \text{with} \quad \hat{q}_t = D(p_t, x_t),$$

and, when realized y_t is observed, to apply a small correction so that δs_t aligns with y_t on average. This keeps the pipeline state consistent even when demand is misspecified or when operational events (e.g., carrier delays) perturb return timing.

Learning also benefits from controlled exploration. Because the platform and stability constraints restrict price movement, we can inject exploration by randomizing within the admissible band (e.g., ± one rung on $\mathcal{P}$ when feasible) with small probability, while still respecting $|p_t - p_{t-1}| \le \Delta$. Such exploration improves identification of both D and ρ and can be scheduled to low-risk periods or low-volume segments.

Nonstationarity is the norm in return behavior (policy changes, seasonality, product mix). We therefore recommend periodic re-estimation of (D, ρ, δ) and, crucially, *re-solving* the outer loop when monitoring detects drift. Because the multipliers are interpretable, practitioners can often diagnose the source of drift: a sustained increase in the learned λ_R signals that the system is spending more return capacity per unit of revenue, while a spike in λ_E indicates tighter effective carbon conditions (higher τ, higher κ, or higher volume pressure through D).

Constraint auditing and "safe" operation. Long-run average constraints are attractive theoretically but must be audited in finite time. We propose an auditing layer that (i) reports rolling averages of returns and emissions, (ii) constructs uncertainty intervals accounting for measurement error in E_t

and stochastic variability in y_t, and (iii) enforces a conservative buffer. Concretely, we track

$$\widehat{R}_T = \frac{1}{T}\sum_{t=0}^{T-1} y_t, \qquad \widehat{E}_T = \frac{1}{T}\sum_{t=0}^{T-1} E_t,$$

and compare them to $\bar{R} - \epsilon_R$ and $\bar{E} - \epsilon_E$, where ϵ reflects desired risk tolerance and statistical uncertainty. If $\widehat{E}_T$ approaches $\bar{E}$ (or the upper confidence bound crosses $\bar{E}$), we temporarily increase λ_E and re-solve the inner loop (or, in the affine approximation, shift the target upward before projection). This "guardrail" is particularly important because the stability constraint can slow adjustment: when Δ is tight, we may need earlier intervention to avoid overshooting a cap during a demand surge.

Two additional checks are operationally valuable. First, we decompose emissions into outbound and returns components, $E_t^S = \kappa_S q_t$ and $E_t^R = \kappa_R y_t$, to confirm whether exceedances are driven by volume (demand shocks) or by reverse logistics (changes in ρ or δ). Second, we log the frequency and magnitude of projection events (how often $\hat{p}_\lambda$ is clipped/rounded) as an indicator of whether platform constraints are binding enough to impede constraint satisfaction; persistent binding suggests either (i) revisiting ladder design, (ii) using additional levers (e.g., shipping/returns policies), or (iii) tightening the buffer and accepting a lower-profit region of the frontier.

Taken together, estimation from logs, the outer–inner loop controller, and an explicit auditing layer translate the theoretical CMDP structure into a deployable pricing system that is both interpretable (via λ as shadow prices) and verifiably compliant with operational and environmental caps.

Numerical illustration (calibration and semi-synthetic evaluation). We close the empirical loop by calibrating a stylized environment to operational magnitudes and then simulating long-run performance under (i) our multiplier-based constrained controller, (ii) a profit-only reinforcement-learning (RL) baseline that ignores the long-run caps, and (iii) a static operations-research (OR) baseline that optimizes a one-period objective with no intertemporal state. The goal of this exercise is not to claim realism of any single calibration, but to stress-test the *mechanisms* emphasized by the theory: (a) delayed returns make the pricing problem genuinely dynamic through the pipeline state s_t; (b) emissions and return constraints generate shadow prices that act like state-dependent marginal costs; and (c) platform stability constraints force gradual adjustment, so robustness matters when primitives drift.

Environment and calibration targets. We simulate weekly periods with state $z_t = (x_t, s_t, p_{t-1})$. The context x_t is a compact vector capturing seasonality and demand shifters; operationally, we treat x_t as a discrete

Markov chain with K regimes estimated from data via clustering on covariates (e.g., traffic and merchandising intensity) and a transition matrix $\widehat{P}$ fit from regime sequences. Conditional on x, demand follows a concave, downward-sloping curve; for transparency we use the linear benchmark

$$D(p, x) = a(x) - bp,$$

where $a(x)$ is chosen to match regime-specific mean weekly unit volume at typical prices, and b is set to match an empirically plausible own-price elasticity around the historical price. Returns are generated by the geometric-delay pipeline described earlier: for each sold unit, eventual return probability is $\rho(p, x)$, and conditional on returning, the arrival delay is geometric with hazard δ. We calibrate δ to match the empirical mean delay (e.g., $\delta = 1/3$ for an average three-week delay), and we fit $\rho(p, x)$ from order-level labels using a logistic specification with a price term, then clip to $[0, 1]$ in simulation. Per-unit costs (c, f, h) are chosen so that (i) gross margin before returns resembles the category's observed margin and (ii) return processing is economically meaningful (i.e., h is nontrivial relative to margin). Emissions factors (κ_S, κ_R) are set so that outbound emissions dominate per-unit carbon impact but returns contribute materially through reverse logistics. Finally, we impose a price ladder $\mathcal{P}$ (e.g., \$1 rungs) and a stability bound Δ (e.g., \$2 per week), reflecting common platform guardrails.

Policies compared. We compare three controllers, all operating under the same admissible action set $\mathcal{A}(z) = \mathcal{P} \cap [p_{t-1} - \Delta, p_{t-1} + \Delta]$.

1. *Constrained dynamic pricing (ours).* We implement the outer–inner loop described above. For a given $\lambda = (\lambda_R, \lambda_E)$, the inner loop solves the discounted MDP with reward $\tilde{\pi}_\lambda = \pi - \lambda_R(y - \bar{R}) - \lambda_E(E - \bar{E})$, using value iteration on a discretized grid for s and the finite regimes for x. The deployed action is the maximizing rung in $\mathcal{A}(z)$. The outer loop updates λ using simulated stationary averages with a small buffer (ϵ_R, ϵ_E) for finite-horizon auditing.

2. *Profit-only RL.* We train an RL agent on the same simulated environment but with reward equal to one-period profit π_t and no penalties for returns or emissions. To keep the comparison focused on the economic difference (constraints versus no constraints), we use a stable policy-gradient method with the same state inputs and the same action constraints $\mathcal{A}(z)$. This baseline captures the common practice of optimizing revenue/profit subject only to platform constraints, and then *checking* externalities ex post.

3. *Static OR (myopic) benchmark.* Each period, this policy chooses

$$p_t \in \arg\max_{p \in \mathcal{A}(z_t)} \pi(p; x_t, s_t),$$

treating the pipeline s_t as affecting current refunds and handling only (through $y_t = \delta s_t$), but ignoring how today's price affects the future pipeline through $\rho(p_t, x_t)D(p_t, x_t)$. This captures a common static margin-optimization logic and clarifies what is lost when we ignore the intertemporal link created by delayed returns.

Evaluation protocol and metrics. We evaluate each policy on long rollouts (e.g., $T = 50{,}000$ periods after burn-in) and report: (i) discounted profit and long-run average profit; (ii) average return arrivals $\bar{y} = (1/T)\sum_{t<T} y_t$; (iii) average emissions $\bar{E} = (1/T)\sum_{t<T} E_t$; (iv) the frequency of constraint violations relative to caps under finite-window audits (rolling windows of length W); and (v) the frequency of binding price projections (how often the chosen action hits $p_{t-1} \pm \Delta$ or the ladder boundary). We emphasize the last two because they diagnose *why* a policy fails: persistent cap exceedances under stability constraints often coincide with frequent boundary hits, indicating insufficient control authority to react to shocks.

Baseline findings in the stationary calibrated environment. In the stationary calibration (where the data-generating primitives match the estimated primitives used by the controller), the outer–inner loop reliably finds multipliers λ such that both long-run average constraints are met with slack roughly equal to the chosen buffers. Economically, the induced pricing rule is intuitive: in high-demand regimes (high $a(x)$), prices rise to ration both shipping volume and the future return pipeline; when the pipeline s_t is elevated, prices rise further because near-term return arrivals $y_t = \delta s_t$ increase effective marginal cost. The profit-only RL policy typically earns the highest raw profit *conditional on ignoring caps* but violates the return and/or emissions constraints substantially; in our calibration the emissions cap is the first to bind when κ_S is large, while the return cap becomes the binding constraint when ρ is high and h is material. The static OR benchmark often respects neither constraint and, even when it happens to satisfy a cap on average, it exhibits higher volatility in resource usage because it ignores how current prices move s_{t+1}. This volatility matters operationally: under finite-window auditing, a policy that meets a cap only in expectation can still be unacceptable if it repeatedly overshoots in short horizons.

Out-of-sample robustness: return-rate shifts. We next introduce a structural break in return behavior meant to mimic a product-mix shift or a policy change (e.g., an extended return window): holding demand fixed, we increase the return propensity by $\Delta\rho > 0$ in all regimes, $\rho^{\text{new}}(p, x) = \min\{1, \rho(p, x) + \Delta\rho\}$. This perturbation is deliberately adverse because it increases both future pipeline inflow and expected reverse-logistics costs. Two patterns are robust across calibrations. First, the profit-only RL policy responds poorly: because it was trained to monetize demand without valuing the shadow costs of returns, it continues to price aggressively, causing s_t to drift upward and generating persistent return-cap violations. Second, our controller remains stable in the sense that the outer loop can re-adjust λ_R upward and restore feasibility with modest profit loss; indeed, the increase in λ_R is a diagnostic statistic that the system is now consuming more return capacity per unit of sales. Importantly, the stability constraint $|p_t - p_{t-1}| \leq \Delta$ creates transient overshoots after the break: even a well-designed controller cannot instantaneously raise prices enough to offset a sudden jump in ρ. This is precisely where the auditing layer and conservative buffers become economically relevant: earlier intervention (raising λ_R preemptively when leading indicators predict a shift) reduces the magnitude and duration of overshoots.

Out-of-sample robustness: carbon-price changes. We also vary the carbon price τ, which in practice can change through regulation, internal carbon accounting, or carrier surcharges passed through as carbon fees. In the simulation, we increase τ unexpectedly and evaluate each policy without retraining its core components, allowing only the minimal operational update that a firm would plausibly implement quickly. The profit-only RL policy again fails the emissions cap because it has no mechanism to internalize the higher carbon cost beyond what is already embedded in π (and if τ is imposed as a *cap* rather than a tax, it does not enter π at all). Our controller adapts cleanly because τ enters the per-period emissions charge τE_t and the cap through the λ_E-update: an increase in τ mechanically raises the effective marginal cost of shipments and returns, while a tighter effective cap raises λ_E. In both cases the model predicts (and the simulation confirms) a monotone shift toward higher prices and lower volume, with the adjustment speed limited by Δ. Practically, this experiment motivates implementing τ and κ as first-class inputs to the controller (not hard-coded constants), so compliance can be maintained without re-estimating demand.

What the numerics add beyond the theory. The theoretical results tell us that the frontier is low-dimensional and that multipliers act as shadow costs; the numerical exercise shows how this plays out under platform frictions and misspecification. Three empirical lessons stand out. (i) *State matters*: policies that ignore the pipeline state s_t can look competitive in average profit but generate large compliance risk under realistic auditing windows. (ii) *Two multipliers are operationally sufficient*: a simple two-dimensional outer loop can correct for sizable shifts in returns or carbon conditions without redesigning the inner-loop solver. (iii) *Stability constraints amplify the value of robustness*: when Δ is tight, early-warning signals and buffers are not conservative bureaucracy; they are economically necessary because the controller's feasible action set limits how quickly the system can re-enter a safe region after shocks.

Summary. Overall, the calibrated simulations support the practical premise of the model: treating returns capacity and emissions capacity as scarce resources with shadow prices yields a pricing policy that is nearly as implementable as profit-only heuristics, but materially more reliable under long-run constraints and under the kinds of nonstationarity that retailers face in the field.

Extensions: coupling, competition, and heterogeneity (and where numerics enter). The baseline model deliberately isolates a single decision margin—a posted price under platform admissibility and stability constraints—so that the economic role of delayed returns and resource shadow costs is transparent. In many retail settings, however, three features matter in tandem: (i) multiple SKUs sharing capacity and congestion, (ii) marketplace competition in which other sellers also price dynamically, and (iii) heterogeneous customers with systematically different demand and return behavior. We sketch each extension in a way that preserves the core logic (Markov sufficiency under geometric delays; low-dimensional multiplier parameterization of constraints), while being explicit about when closed-form structure breaks and numerical methods become essential.

(i) Multi-SKU coupling through shared congestion and shared caps. Let SKUs be indexed by $i \in \{1, \dots, n\}$. Each SKU has price $p_{i,t}$, demand $q_{i,t} = D_i(p_{i,t}, x_t)$, and a return pipeline $s_{i,t}$ evolving (under the same geometric-delay logic) as

$$s_{i,t+1} = (1-\delta)s_{i,t} + \rho_i(p_{i,t}, x_t)D_i(p_{i,t}, x_t), \qquad y_{i,t} = \delta s_{i,t}.$$

The most operationally relevant coupling comes from shared fulfillment capacity and carrier surcharges that depend on total outbound volume $Q_t = \sum_i q_{i,t}$. A natural congestion cost is then $G(Q_t)$, convex and increasing. Likewise, platform- or firm-level constraints often apply in aggregate, e.g.,

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t<T}\sum_i y_{i,t} \le \bar{R}, \qquad \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t<T}\sum_i E_{i,t} \le \bar{E},$$

with $E_{i,t} = \kappa_S q_{i,t} + \kappa_R y_{i,t}$. Two implications follow. First, the Markov state expands to $(x_t, s_{1,t}, \dots, s_{n,t}, p_{1,t-1}, \dots, p_{n,t-1})$, so even though each pipeline remains one-dimensional, the overall state is $O(n)$. Second, separability across SKUs is broken by $G(\sum_i q_{i,t})$: the marginal congestion cost $G'(Q_t)$ acts like a common endogenous surcharge that depends on the joint pricing vector.

Conceptually, the Lagrangian relaxation still yields a small number of *global* shadow prices (λ_R, λ_E) for the return and emissions caps, so the Pareto frontier in $(\text{profit}, \text{returns}, \text{emissions})$ remains parameterized by at most two multipliers under the same convexity/Slater conditions. What changes is the *inner* optimization: the Bellman maximization becomes a coupled choice of $p_t = (p_{1,t}, \dots, p_{n,t})$ over a Cartesian product of ladder-and-stability sets. In special cases, one can partially recover structure by introducing an auxiliary "congestion price" $\gamma_t \approx G'(Q_t)$ and solving per-SKU best responses conditional on γ_t, iterating to consistency (a form of dual decomposition). But in general, and especially when $\mathcal{P}$ is discrete and stability constraints bind, closed-form affine rules like Proposition 3 do not survive: numerical dynamic programming, approximate value function methods, or policy optimization become necessary to compute implementable policies for moderate n.

(ii) Multi-agent marketplace competition and strategic feedback. In a marketplace, a single retailer's demand depends not only on its own price but on competitors' prices and availability. A reduced-form way to incorporate this is to embed competitor conditions in the context state x_t, for example by letting

$$q_t = D(p_t, x_t), \quad x_t = (\text{seasonality}, \text{traffic}, \text{competitor index}, \dots),$$

where the competitor index evolves exogenously via a Markov kernel $P(\cdot \mid x)$ estimated from observed marketplace dynamics. Under this interpretation, our framework remains a single-agent constrained MDP: competitors are part of the environment. This is empirically convenient and often defensible when the retailer is small relative to the marketplace or when competitors' pricing is noisy and not tightly coupled to any one seller.

If instead we treat competitors as strategic agents who also solve dynamic pricing problems (possibly with their own returns and emissions constraints), the correct object is a constrained Markov game with platform restrictions. In such a setting, the analogue of our multiplier-based approach can be used in at least two ways. One is *agent-level*: each seller solves a Lagrangian-relaxed control problem taking rivals' policies as given, yielding a best-response mapping in multipliers and policies, and we compute a Markov perfect equilibrium numerically. The second is *platform-level*: the platform chooses (or induces) shadow prices for system-wide caps (e.g., emissions), effectively implementing a Pigouvian tax or quota price that all sellers face, while allowing decentralized best responses. The latter perspective is attractive because it connects directly to policy instruments, but it also highlights a limitation: equilibrium existence and uniqueness are no longer guaranteed by the convex CMDP logic alone, because strategic interactions can introduce non-convexities in the induced performance set and discontinuities when action sets are discrete (price ladders) and stability constraints bind. For this reason, even when the economic mechanism is clear (shadow costs push equilibrium prices upward and volumes downward), equilibrium computation is typically numerical.

(iii) Heterogeneous customer segments and return heterogeneity. Customer heterogeneity matters because return propensity is not merely a function of price and product; it varies systematically across segments (e.g., new versus repeat customers, size-sensitive shoppers, or geographically distinct shipping regions). A parsimonious extension introduces segments $g \in \{1, \dots, G\}$ with segment-specific demand $D_g(p, x)$ and return propensity $\rho_g(p, x)$. If the retailer posts a single price p_t, aggregate demand is

$$D(p, x) = \sum_{g=1}^{G} D_g(p, x),$$

and under *common geometric delay* δ, the aggregate pipeline remains one-dimensional:

$$s_{t+1} = (1 - \delta)s_t + \sum_{g=1}^{G} \rho_g(p_t, x_t) D_g(p_t, x_t), \qquad y_t = \delta s_t.$$

In this case, our Markov reduction and multiplier interpretation go through essentially unchanged; segment heterogeneity is "compressed" into the mapping $p \mapsto \sum_g \rho_g(p, x) D_g(p, x)$. Economically, this is useful: it tells us that a single scalar pipeline state can remain sufficient even with rich cross-sectional heterogeneity, so long as return timing is memoryless and costs scale linearly with total returns.

Where numerics (and higher-dimensional state) enter is when heterogeneity affects *return timing* or *costs* in a way that breaks aggregation. If segments have different hazards δ_g (e.g., because some customers are systematically slower to return), then the correct sufficient statistic becomes a vector $s_t = (s_{1,t}, \dots, s_{G,t})$ with

$$s_{g,t+1} = (1 - \delta_g)s_{g,t} + \rho_g(p_t, x_t) D_g(p_t, x_t), \quad y_t = \sum_g \delta_g s_{g,t},$$

and pricing must manage a genuinely multidimensional pipeline. Similarly, if emissions factors or handling costs vary by segment (e.g., remote regions with higher κ_S and κ_R), then even with a common δ, the per-period objective depends on the segment composition of shipments and returns, again pushing us toward vector-valued states or richer context variables. In these cases, projected affine rules become approximations rather than exact solutions, and one typically relies on discretization, approximate dynamic programming, or function approximation (e.g., linear or neural value functions) with the same multiplier outer loop enforcing long-run feasibility.

What persists across extensions. Across all three directions, two pieces of economic structure remain robust. First, delayed returns create an intertemporal externality that is summarized by a small state *per independent return-timing process* (scalar under geometric delay, vector under multiple hazards). Second, long-run caps on returns and emissions remain naturally priced by a small set of multipliers (λ_R, λ_E), which can be adjusted in an outer loop using observed stationary averages even when the inner control problem is high-dimensional. What changes is not the logic of shadow costs, but the computational burden of mapping state to an action under coupling, strategic feedback, and heterogeneous dynamics—precisely the cases in which numerics are not an optional embellishment, but the method by which the theory becomes operational.

Policy and platform implications: effective constraints, carbon pricing, and auditable compliance. Our framework is useful precisely because it separates *what* society or the platform wants to limit (returns handling burden and shipping-related emissions) from *how* the retailer chooses to respond (a dynamically adjusted posted price, subject to admissibility and stability). The central practical message is that long-run caps and per-unit carbon fees enter the retailer's problem through a small set of shadow costs. This is more than a mathematical convenience: it provides a concrete design principle for policy and platform rules. If the platform can translate abstract objectives ("reduce returns" or "reduce emissions") into stable, predictable marginal incentives, then decentralized pricing decisions will internalize the external costs without requiring the platform to solve the retailer's entire dynamic program.

Which constraints are most effective? Target the bottleneck, not the symptom. A return-volume cap $\bar{R}$ and an emissions cap $\bar{E}$ both reduce shipped volume in equilibrium, but they do so through distinct channels. A cap on returns directly prices the scarce resource in reverse logistics (warehouse labor, inspection capacity, disposal constraints), and is therefore most effective when the operational system is genuinely constrained by return handling. In contrast, an emissions cap (or equivalently a binding emissions price) targets the environmental externality; it is most effective when the social objective is to reduce total shipping footprint rather than to protect internal processing capacity. In many settings the two objectives are aligned, but the alignment is not mechanical: a retailer can reduce $\mathbb{E}[y_t]$ by shifting sales toward lower-return variants or better-described products without proportionally reducing outbound shipments, whereas an emissions constraint pushes on both outbound and returns via $E_t = \kappa_S q_t + \kappa_R y_t$. The model therefore suggests a "bottleneck test": if the main harm is operational congestion in returns, a return-cap instrument (or a return-processing fee) is more direct; if the harm is the environmental footprint, the emissions instrument is the right target.

Fees versus caps: predictability versus hard guarantees. From an implementation perspective, per-unit fees and hard caps differ in how they trade off predictability and guarantees. A carbon fee τ provides a stable marginal signal and is easy to administer; it does not guarantee hitting a specific $\bar{E}$, but it avoids the discontinuities that arise when hard caps bind. Conversely, a cap provides a hard quantity guarantee but requires either (i) rationing or (ii) a shadow price λ_E that can move over time as scarcity fluctuates. Our Lagrangian formulation reconciles these: when $\bar{E}$ is enforced as a long-run average, the optimal policy behaves *as if* there were an endogenous carbon surcharge $\lambda_E \tau$ that is adjusted until measured stationary emissions meet the cap. This suggests a practical "outer-loop" platform design: rather than policing every price, the platform can update category-level fees (interpretable as λ_E and λ_R) based on observed rolling averages of emissions and returns, while allowing sellers to optimize locally.

How carbon charges shift optimal prices (and why returns amplify pass-through). In the one-period objective, carbon pricing enters as $-\tau(\kappa_S q_t + \kappa_R y_t)$, which behaves like an increase in marginal cost of outbound shipments and return arrivals. Under downward-sloping demand, the comparative static is typically $\partial p^* / \partial \tau > 0$, but the magnitude depends on both demand elasticity and the returns pipeline. Intuitively, when returns are substantial, a marginal increase in sales creates not only current outbound emissions but also future return emissions and processing costs. This raises the *dynamic* marginal cost of selling an extra unit. Even if the immediate carbon charge is small, the discounted stream of expected return-related charges can be meaningful when $\rho(p, x)$ is high or when δ is large (fast returns make the costs arrive sooner). As a result, categories with high return propensity should exhibit larger effective pass-through of carbon fees into prices, all else equal, because the same outbound sale carries more downstream carbon liability through s_{t+1}.

Interaction with stability constraints: why rigid pricing rules can undermine environmental objectives. Platform stability constraints $|p_t - p_{t-1}| \leq \Delta$ are typically motivated by consumer trust and by avoidance of "price gouging" perceptions. Our analysis highlights a less discussed effect: tight Δ can make environmental and returns constraints harder (and costlier) to satisfy, because the retailer cannot respond quickly to shocks in x_t (e.g., demand surges) or in the pipeline s_t (e.g., elevated pending returns). When Δ binds, the policy effectively becomes a clipped version of the unconstrained optimum, so the adjustment to a higher τ or a tighter $\bar{E}$ may be delayed. This does not mean stability constraints are undesirable; rather, it suggests that if the platform insists on tight Δ, it should anticipate the need for *stronger* shadow costs (larger λ_E or τ) to achieve the same emissions outcomes, or should complement stability with non-price levers (better product information, sizing tools, or shipping-mode changes) that reduce κ_S, κ_R and $\rho(p, x)$ directly.

Welfare: separating private profit, consumer surplus, and external costs. The objective we solve is retailer profit subject to constraints, which is the right positive model for platform compliance but not the full welfare criterion. For welfare analysis, we would add consumer surplus and subtract external damages from emissions and waste. Carbon pricing has a canonical welfare interpretation: when τ reflects the social cost of carbon (and when emissions measurement is accurate), the induced price increase can be welfare-improving even if it reduces consumer surplus, because it internalizes external harm. Returns complicate this logic because lenient returns generate consumer option value (insurance against misfit) while also creating processing, congestion, and environmental costs. A return cap $\bar{R}$ can therefore raise welfare when reverse-logistics costs are largely social (waste, landfill constraints, transport emissions), but it may reduce welfare if it is implemented bluntly (e.g., by discouraging legitimate purchases from uncertain consumers) rather than by improving match quality. The model's main welfare lesson is to distinguish *volume reduction* from *match improvement*: policies that reduce $\rho(p, x)$ (better information, sizing, packaging) can dominate policies that merely raise prices to shrink q_t.

Instrument design: what platforms can implement without solving the seller's problem. Because the shadow-cost logic is additive, platforms can implement approximate compliance with limited information by posting *linear* surcharges: a fee per outbound unit proportional to κ_S and a fee per return proportional to κ_R, with an additional return-handling fee capturing congestion in processing. In our notation, this corresponds to operationalizing $\lambda_E \tau \kappa_S$ and $\lambda_E \tau \kappa_R + \lambda_R$ as explicit charges. Importantly, these need not be uniform across products: if the platform can estimate SKU-level or category-level κ factors and average ρ, then differentiated fees align incentives with heterogeneous footprints. Compared with direct price controls, such fees preserve seller autonomy and are robust to the platform's limited ability to observe $D(\cdot)$ or to forecast x_t.

Auditing and measurement: what must be verified for the model's prescriptions to be credible. Any policy that relies on emissions accounting or return caps is only as good as its measurement system. Three components are critical. First, outbound shipments q_t and return arrivals

y_t must be measured consistently, including cancellations and multi-item orders, because mismeasurement can distort both the estimated pipeline s_t and the assessed charges. Second, emissions factors κ_S, κ_R must be auditable and periodically updated: shipment distance, carrier mix, packaging weight, and consolidation practices change over time, so static factors invite either drift or gaming. Third, platforms should audit the *return timing process* to validate the geometric-delay approximation in any compliance-critical application. Our Markov reduction is exact under geometric delays; when delays are duration-dependent, a single scalar s_t may understate near-term return risk, and a seller could appear compliant in expectation while generating clustered operational stress.

Auditing recommendations: rolling windows, category benchmarks, and gaming-resistant metrics. Operationally, we recommend a three-layer audit. (i) *Rolling-window reporting*: compute realized averages of y_t and E_t over overlapping windows to detect persistent exceedances early, consistent with the long-run nature of $\bar{R}$ and $\bar{E}$. (ii) *Category benchmarks*: compare realized return rates and emissions intensities to peer baselines conditional on observable x_t (seasonality, region), which helps separate genuine product problems from demand shocks. (iii) *Gaming-resistant accounting*: ensure that incentives do not encourage relabeling returns as exchanges, splitting shipments to manipulate κ factors, or shifting fulfillment off-platform. In our language, the goal is to make the measured q_t, y_t, E_t close to the true physical quantities so that the shadow costs λ_R, λ_E guide behavior rather than guide accounting choices.

Limitations and a practical synthesis. We emphasize two limitations. First, price is not the only lever that affects emissions and returns; improving fulfillment operations can lower κ_S, κ_R, and product design and information can lower $\rho(p, x)$. Second, stability and ladder constraints create discrete, sometimes non-smooth decision rules that can weaken first-order comparative statics in finite samples. Nonetheless, the policy synthesis is clear: (a) choose instruments that price the true scarce resources (emissions capacity and return-processing capacity), (b) implement them via auditable marginal charges or dynamically updated shadow fees, and (c) recognize that rigid price constraints may require complementary non-price interventions to achieve environmental and operational targets at low welfare cost.

7 Conclusion

We study a dynamic pricing problem in which a retailer faces two operational externalities that are increasingly central in practice: product returns and shipping-related emissions. The distinctive modeling challenge is that both externalities are intrinsically intertemporal. Returns arrive with delay, creating a pipeline of future reverse-logistics workload and refund exposure; emissions are generated by both outbound shipments and subsequent returns. At the same time, real platforms often restrict how the retailer may adjust prices through admissibility ladders and explicit stability rules. Our goal is therefore not only to characterize an optimal policy in principle, but also to identify a compact state representation and an implementable policy structure that can be audited and tuned with a small number of interpretable parameters.

The first contribution is a Markov reduction of delayed returns. Under geometric return delays, the expected flow of future return arrivals can be summarized by a scalar "pipeline" state s_t that evolves as

$$s_{t+1} = (1 - \delta)s_t + \rho(p_t, x_t)D(p_t, x_t),$$

with current expected return arrivals $y_t = \delta s_t$. This reduction is not merely a technical convenience: it clarifies what a retailer must track to manage returns dynamically. Rather than remembering the full history of sales by cohort, the retailer only needs the current mass of pending returns, which aggregates past sales through the memoryless property. In our setting, the relevant decision state becomes (x_t, s_t, p_{t-1}): covariates that shift demand and return propensity, the return pipeline that governs near-term arrivals, and the previous price that determines the stability-feasible set.

The second contribution is a low-dimensional description of the performance frontier when the retailer faces long-run average constraints. We impose caps on mean return arrivals and mean emissions, motivated by operational capacity in reverse logistics and by environmental policy. While such constraints appear to create a complex intertemporal feasibility problem, standard convexity and Slater-type conditions imply that every Pareto-efficient operating point—in terms of profit, mean returns, and mean emissions—can be implemented by solving an unconstrained dynamic program with a two-dimensional vector of multipliers (λ_R, λ_E). In this sense, the efficient set is parameterized by at most two scalars. This perspective matters for implementation: it suggests that a platform or regulator need not directly control the retailer's entire policy. Instead, it can adjust a small number of shadow prices (interpretable as fees) until measured long-run averages meet operational or environmental targets.

The third contribution is structural: in a linear demand and affine return-propensity specification, the optimal price rule in the Lagrangian-relaxed problem is approximately affine in the key state variables and shadow costs, and the platform and stability restrictions enter through a projection. Concretely, the unconstrained optimizer $\hat{p}(x_t, s_t; \lambda_R, \lambda_E)$ is clipped to the feasible interval

$$p_t^* = \Pi_{\mathcal{P}\cap[p_{t-1}-\Delta,\, p_{t-1}+\Delta]}\big(\hat{p}(x_t, s_t; \lambda_R, \lambda_E)\big).$$

This “projected affine” structure yields a practical algorithmic template: estimate local demand and return responses, compute an unconstrained target price that internalizes carbon and return shadow costs, and then apply transparent platform-compliance rules. The projection form also makes clear when and why stability rules bind: the retailer would like to move the price more aggressively in response to shocks in x_t or s_t, but the platform restricts the step size.

Across these results, a unifying economic message emerges: returns and emissions constraints enter the pricing problem like additional marginal costs, and their dynamic consequences are mediated by the return pipeline. In the per-period objective, emissions pricing penalizes both outbound and return shipments; return constraints penalize the arrival flow. In the continuation value, today’s price affects tomorrow’s costs by changing s_{t+1} through $\rho(p_t, x_t)D(p_t, x_t)$. As a result, the incentive to raise price in high-return states is not simply about lower expected net revenue today, but about reducing the stock of future reverse-logistics pressure. This helps interpret why seemingly similar categories can exhibit different price dynamics: two products with identical demand elasticities but different return propensities ρ or different return-speed parameters δ will have different dynamic marginal costs of selling an extra unit.

We also view the framework as a disciplined way to connect platform rules to operational realities. Platforms often adopt price ladders and stability constraints to promote transparency and consumer trust. Our analysis highlights the tradeoff: such constraints can be welfare-improving on their own terms, but they reduce the retailer’s ability to respond to shocks that are relevant for emissions and returns. The model does not argue against stability; rather, it makes precise how stability restrictions can necessitate higher shadow costs (or complementary non-price interventions) to achieve the same environmental or reverse-logistics outcomes. The projection characterization is useful here because it provides a direct diagnostic: frequent clipping events indicate that the platform is constraining the retailer’s primary adjustment margin, so the platform should expect either higher compliance costs or greater reliance on alternative levers.

Several limitations point to natural extensions. First, the geometric-delay assumption is essential for the one-dimensional pipeline state. When return delays are duration-dependent or depend on cohort characteristics (e.g., holiday gifting, carrier disruptions), the sufficient state becomes higher-dimensional, and the computational burden grows. Second, we focus on price as the principal control, but in many environments the retailer can also affect $\rho(p, x)$, κ_S, and κ_R through information provision, packaging, consolidation, and shipping mode. Endogenizing these operational levers would enrich the set of feasible tradeoffs and may reduce the need for blunt volume reductions via higher prices. Third, our baseline presentation abstracts from strategic consumer behavior (e.g., “bracketing” and anticipatory returns). Incorporating forward-looking consumers could amplify the value of stability constraints for trust, while also changing the mapping from price to returns.

On the empirical and computational side, the model suggests a concrete agenda. Because the policy-relevant objects are elasticities and emissions intensities, the platform can prioritize measurement of $D(p, x)$, $\rho(p, x)$, and shipment-return emissions factors. The multiplier parameterization then provides a simple outer-loop calibration: adjust (λ_R, λ_E) based on observed rolling averages until caps are met, and let the retailer's inner-loop optimization manage day-to-day pricing within platform constraints. Even when closed forms are unavailable (e.g., with general $G(\cdot)$ or richer state dynamics), the same structure supports approximate dynamic programming: the state is compact, the multipliers are low-dimensional, and the projection step is transparent.

We close with a synthesis. The model is designed to illuminate a specific tradeoff that practitioners face: operational and environmental constraints are real, but so are platform rules restricting price dynamics. By showing that delayed returns admit a low-dimensional state, that constrained objectives admit a low-dimensional shadow-price representation, and that platform rules often act through simple projection, we obtain both theory and a blueprint for implementation. The broader lesson is that effective policy and platform design should focus on pricing the true scarce resources—reverse-logistics capacity and emissions capacity—while preserving enough flexibility for retailers to adapt to evolving demand and return conditions.