DECISION-TABPFN: SEPARATING ACTION SCORING FROM INVENTORY CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

A frozen predictor can select low-loss actions on fixed queries yet yield a weaker sequential policy. We study this gap in perishable inventory using Decision-TabPFN, which scores immediate profit from completed transitions. Matched comparisons show a TabPFN advantage over XGBoost on Decision-TabPFN histories, but this advantage varies with the source of queried states. Holding the backbone fixed, independently selected forecast-to-order pipelines earn more profit in two of three regimes. Fixed-budget context comparisons show that offline rankings need not persist online. To connect prediction with control, we bound inventory-dependent continuation value independently of the remaining horizon and cancel forecasting error shared by policies. The resulting inequality separates demand-prediction error from model-planning error and yields sublinear dynamic-optimum regret under explicit cyclic-demand learning and complete-horizon planning. A separate block-planning study exposes strong prior sensitivity and no demonstrated benefit from TabPFN proposals. Together, the results distinguish action scores on deployment states, the supplied decision structure, and continuation value as separate evaluation targets.

1 INTRODUCTION

Can a strong frozen predictor become a strong inventory controller simply by scoring possible orders? This interface is attractive because it needs no policy training: completed state–action–profit transitions become regression rows, and candidate orders become queries. Perishable inventory exposes its central difficulty. An order determines current profit, tomorrow's age profile, and the history from which later actions are predicted. A local predictive comparison therefore measures only part of the decision problem.

We study this question through **Decision-TabPFN**, a frozen TabPFN regressor that scores immediate profit and samples among high-scoring orders. The policy adapts through operational history without receiving demand-period or phase labels. Frozen tabular action selection already appears in contextual bandits (Gupta et al., 2026); our focus is the additional structure needed when actions carry inventory into future periods. Direct scoring provides a test case for this transition from prediction to sequential control.

The empirical obstacle is attribution. Higher policy profit can reflect a better predictor, a stronger ordering rule, or a different distribution of visited states. We compare predictors on identical histories and queries, compare decision interfaces using the same TabPFN backbone, and compare contexts under a fixed row budget. The resulting picture is specific: the scorer improves action selection on its own policy's histories, yet forecasting pipelines that supply inventory structure can outperform direct scoring. The advantage also depends on the source of the queried states. Context rankings change between fixed-query evaluation and online use, where both the history and the action rule differ.

The theoretical obstacle is continuation value. Even exact immediate-profit predictions omit the value of stock carried forward. A generic prediction-error bound multiplied by the entire remaining horizon obscures the inventory structure. We instead bound the value of initial stock using its finite lifetime and a common-order coupling. Centering continuation values at empty stock cancels the forecasting error shared by all policies. This gives a prediction-to-control inequality with separate terms for demand error and planning error, followed by an explicit unknown-period regret

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

guarantee. The guarantee requires complete-horizon planning; the separately evaluated two-period implementation remains sensitive to the prior and shows no TabPFN-specific gain.

Our contributions are: **(i)** controlled evidence locating the gap between predictor quality and decision-interface performance; **(ii)** source-stratified and fixed-budget context comparisons that delimit what local scoring results imply online; and **(iii)** an inventory-value argument that converts exogenous prediction error and verified planning error into dynamic-optimum regret.

2 Related Work

Frozen models for decisions. Prior-data fitted networks amortize prediction through pretraining (Müller et al., 2022; Hollmann et al., 2023; 2025). BC-ICL combines frozen tabular action scores with bootstrap randomization for contextual bandits (Gupta et al., 2026). ICGPS addresses censored repeated-newsvendor demand through learned generative completions (Mukherjee et al., 2026). Here, stock carries over and expires, while shortage-cost feedback reveals demand. These differences motivate our interface and continuation analysis, without making either preceding approach a direct baseline. REGENT learns a retrieval-augmented policy (Sridhar et al., 2025); our context experiments concern a fixed tabular checkpoint and a joint retrieval rule.

Inventory and exogenous control. Classical inventory policies exploit cost and demand structure (Scarf, 1960; Porteus, 2002; Huh et al., 2009); perishability adds age-dependent decisions (Nahmias, 1982; Karaesmen et al., 2011; Chao et al., 2015; Zhang et al., 2018). Learning methods include deep reinforcement learning and gradient–bandit policies (Gijsbrechts et al., 2022; Yuan et al., 2021). Gong & Simchi-Levi (2024) establish square-root-order regret for nondiscarding cyclic-demand lost sales; Chen et al. (2024) use coupling to learn under censored demand and uncertain supply. Exogenous-MDP methods exploit known dynamics and external inputs (Sinclair et al., 2023; Wan et al., 2026). Our contribution is a horizon-independent inventory-value bound and shared-history cancellation for continuing perishable control. Section 6.3 and Appendix A.13 distinguish the assumptions and proof mechanisms.

Predictive evaluation under feedback. Sequential prediction must account for policy-induced state distributions (Ross et al., 2011). We apply this distinction at identical inventory queries and across independently generated policy histories. Theoretical constructions for in-context learning need not describe a trained checkpoint (von Oswald et al., 2023); our guarantee accordingly concerns an explicit demand model and a checked planner. Independent validation of the forecasting pipeline follows the broader need to control tuning when comparing tabular models (Erickson et al., 2025).

3 Problem Formulation

We consider a single product with lifetime m, zero replenishment lead time, and a finite order set $\mathcal{A}$. At the end of period $t-1$, the residual inventory vector is $\mathbf{v}_{t-1} \in \mathbb{Z}_+^m$, ordered from oldest to youngest, and outstanding backlog is b_{t-1}. Lost-sales systems have $b_{t-1} = 0$. A policy chooses a_t using only the information available before demand in period t.

We use the simulator's accounting convention: at the start of period t, the oldest residual cohort expires, the remaining cohorts shift, and order a_t enters the youngest cohort. Demand is then served under FIFO or LIFO. Let y_t, u_t, and e_t denote units sold, unmet total requirement, and units expiring, respectively. The recorded profit is

$$P_t = ry_t - K\mathbf{1}\{a_t > 0\} - ca_t - h\|\mathbf{v}_t\|_1 - pu_t - ee_t. \tag{1}$$

In backlog systems the requirement is $D_t + b_{t-1}$ and $b_t = u_t$; in lost-sales systems it is D_t and $b_t = 0$. Charging expiration at the start of a period fixes the label used for profit prediction, even though an alternative end-of-period convention can describe the physical aging process. Appendix B gives the transitions.

Synthetic demand is independent across periods with $D_t \sim F_{\phi_t}$, where $\phi_t = 1 + ((t-1) \bmod H)$. The learner is given neither H nor $F_1, \ldots, F_H$. It observes completed sales, age inventories, backlog, and the realized cost components. In particular, exact shortage cost and its known positive unit penalty reveal u_t. This is operational-cost feedback, not a sales-only censored-demand setting.

The objective is $T^{-1}\mathbb{E}[\sum_{t=1}^{T} P_t]$. We compare online policies with hindsight-tuned constant and phase-specific base-stock or (s, S) references. These references use additional information and restricted policy families; their reported profits are not upper bounds on optimal perishable-inventory control.

4 DECISION-TABPFN

The policy learns a local action score from completed transitions. It then applies that score repeatedly in a system whose state depends on past actions. This separates the supervised target from the long-horizon objective.

4.1 FROM OPERATIONAL HISTORY TO REGRESSION ROWS

For the main short-lifetime setting, a completed observation has the ordered features

$$\mathbf{h}_t = (a_t, c_t^o, \mathbf{v}_t, c_t^h, e_t, c_t^e, b_t, c_t^u, y_t, C_t, R_t, P_t), \tag{2}$$

where $c_t^o, c_t^h, c_t^e, c_t^u$ are the four cost components, C_t is total cost, and R_t is revenue. Thus the inventory entries are an age vector, and the shortage-state entry is backlog (zero in lost-sales systems). For $m = 3$, the observation has 14 scalar entries; adding the candidate order gives 15 input features. The long-lifetime reference implementation compresses inventory to its total and omits expiration features, as specified in Appendix B.2.

At the beginning of period t, the in-context data and candidate queries are

$$\begin{aligned} \mathcal{D}_{t-1} &= \{(\mathrm{Cat}(\mathbf{h}_{k-1}, a_k), P_k) : 2 \le k < t\}, \\ X_t^a &= \mathrm{Cat}(\mathbf{h}_{t-1}, a), \qquad \widehat{\mu}_t(a) = q_\theta(X_t^a \mid \mathcal{D}_{t-1}). \end{aligned} \tag{3}$$

The pretrained parameters θ remain fixed; the estimator's fit interface only refreshes the context. Past actions determine the labels and states available for future queries. The implemented AC stores only the executed action's profit, although the feedback permits the additional labels described in Section 4.3.

4.2 FROM ACTION SCORES TO AN ORDER

All candidate actions are scored in a batched prediction call. Let $\mathcal{A}_t^B$ contain the B highest-scoring actions, with deterministic tie-breaking, and let $\bar{\mu}_t$ and s_t be the mean and standard deviation of their scores. We use

$$z_t(a) = \frac{\widehat{\mu}_t(a) - \bar{\mu}_t}{s_t + \epsilon}, \qquad p_t(a) = \frac{\exp(z_t(a)/\tau_t)}{\sum_{a' \in \mathcal{A}_t^B} \exp(z_t(a')/\tau_t)}, \tag{4}$$

$$\tau_t = \max\{\tau_{\min}, \tau_0/(1 + \eta j_t)\}, \tag{5}$$

where j_t counts completed model-based decisions. Top-B selection limits where the policy samples, while standardization makes its temperature less sensitive to the units of profit. Filtering occurs after scoring and therefore does not reduce the number of queried actions. It also provides no exploration guarantee outside the retained set.

The default configuration uses $B = 10$, $\epsilon = 10^{-6}$, $\tau_0 = 2$, $\eta = 0.1$, and $\tau_{\min} = 0.1$. Four initial orders, rounded from independent uniform draws on $[200, 400]$, provide three completed transition rows for the first prediction. The temperature decreases to a positive floor, retaining stochastic choice among the highest-scoring orders. Appendix B.2 provides implementation details.

4.3 CONTROLLED DECISION INTERFACES

AC and FO use the same completed observation $\mathbf{h}_{t-1}$ and cost feedback. AC predicts profit for candidate actions; FO predicts demand and supplies an explicit ordering rule. Neither receives phase labels or a period bound. The theoretical demand mixture additionally uses the clock and $H_{\max}$; the informed oracle knows the true phase laws.

For a pre-decision inventory $\mathbf{v}$, backlog b, and candidate order a, let $S = \sum_{i=2}^{m} v_i + a$ be stock available after expiration. The known one-period profit for demand d is

$$G(\mathbf{v}, b, a, d) = r \min\{S, d + b\} - ca - K\mathbf{1}\{a > 0\} \\ - h(S - d - b)^+ - p(d + b - S)^+ - ev_1. \tag{6}$$

Backlog is zero in lost-sales systems. Structured FO predicts demand quantiles $\widehat{d}_{t,j}$ from $\mathbf{h}_{t-1}$ and chooses

$$a_t^{\mathrm{FO}} = \arg\max_{a \in \mathcal{A}} \frac{1}{J} \sum_{j=1}^{J} G(\mathbf{v}_{t-1}, b_{t-1}, a, \widehat{d}_{t,j}), \qquad J = 129. \tag{7}$$

This rule evaluates all actions through an explicit cost model. AC instead scores unexecuted actions by generalizing from the realized profits in its context. Both optimize an immediate target; only their deployed trajectories incorporate the resulting future states.

What the structural interface supplies. Write $I = \sum_{i=2}^{m} v_i$ and $g_t(a) = \mathbb{E}[G(\mathbf{v}, b, a, D_t) \mid \mathcal{H}_{t-1}]$, where $\mathcal{H}_{t-1}$ is the available history. For integer demand and $0 \le a < A$,

$$g_t(a+1) - g_t(a) = (r + p + h) \Pr(D_t + b > I + a \mid \mathcal{H}_{t-1}) - (c + h) - K\mathbf{1}\{a = 0\}. \tag{8}$$

Thus the positive-order profit curve has nonincreasing increments. Structured FO supplies this shape through a common demand distribution; unconstrained action regression must infer it from its labels. The identity describes a structural difference between the interfaces, not an experimentally isolated cause of their profit gap.

Full feedback also gives $D_t = y_t + c_t^u/p - \mathbf{1}\{\mathrm{BL}\}b_{t-1}$ when $p > 0$. One can therefore construct $G(\mathbf{v}_{t-1}, b_{t-1}, a, D_t)$ for every candidate action at the *visited* state. These labels share one demand observation and describe only the visited state. Our comparison evaluates the implemented one-label AC; counterfactual-label augmentation remains untested.

5 Separating Predictor Quality from Policy Performance

The controlled suite asks three questions: where the scorer helps, whether this benefit survives a change of decision interface, and what remains beyond immediate scoring. The principal regimes are lost sales (LS) with FIFO or LIFO and $K = 0$, and backlog (BL) with LIFO and $K = 20$. Each uses lifetime $m = 3$, orders 0–600, and ten cyclic demand laws. Revenue, unit order, holding, and expiration costs are 10, 3, 1, and 4; shortage penalties are 15 for LS and 3 for BL.

Evaluation. Five validation seeds select FO; ten disjoint test seeds evaluate it. Demand and four initial orders are paired across methods. Profits average all periods, including initialization: $T = 10{,}000$ for the primary AC–FO comparison and $T = 3{,}000$ for context controls. Intervals use paired Student-t uncertainty across seeds. The three primary tests receive Holm correction; auxiliary intervals are pointwise. Full protocols appear in Appendices B–C. The earlier eight-regime benchmark is retained in Appendix F, with its original comparators and protocol.

A trajectory view of the benchmark. Figure 1 shows all eight original benchmark regimes. Adaptation and final rankings vary with the inventory regime and setup cost. The controlled tests below ask what drives this ranking, using separate paired seeds and independently selected FO. Terminal profits, spoilage, and cycle-response diagnostics appear in Appendix F.

5.1 Where does the frozen scorer help?

TabPFN and XGBoost receive identical histories and score the same 601 actions. Twenty late queries per seed cover all phases, separately from native AC, greedy XGB-AC, and structured FO trajectories. Using 2,048 shared phase-conditional demand draws, we estimate expected profits $\bar{\mu}_{s,t}(a)$ and evaluate greedy action loss

$$\ell_{s,t}(f) = \max_{a \in \mathcal{A}} \bar{\mu}_{s,t}(a) - \bar{\mu}_{s,t}\left(\arg\max_{a \in \mathcal{A}} f(a)\right). \tag{9}$$

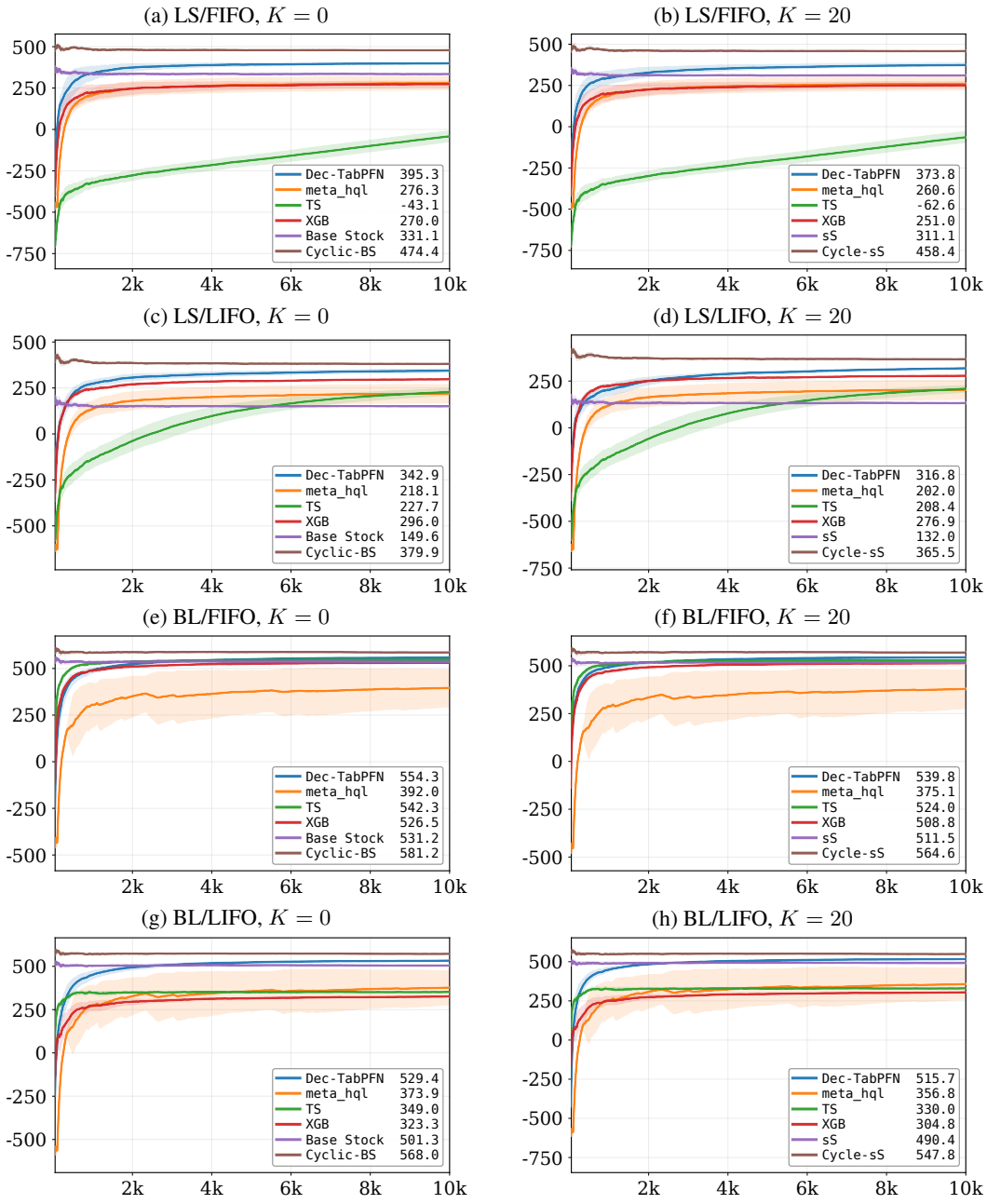

Figure 1: **Profit trajectories across all eight inventory regimes.** Columns distinguish $K = 0$ and $K = 20$; rows distinguish LS/FIFO, LS/LIFO, BL/FIFO, and BL/LIFO. Original benchmark, ten runs over 10,000 periods; cumulative mean profit with pointwise 95% intervals. Legends report full-horizon means; XGB denotes the demand-forecasting baseline. Classical and cyclic curves are hindsight-tuned references, with the true period supplied only to the cyclic references. Appendix F reports terminal profits and complementary diagnostics.

Phase information is used only for this offline target. Queries are averaged within source and seed before uncertainty is computed.

The advantage is conditional on the source distribution. Full-context TabPFN has lower loss on AC-generated histories in all three regimes (Table 1). FO-generated histories support a gain in LS/FIFO, but leave both LIFO comparisons unresolved; the BL/LIFO mean reverses. The prespecified equal-source pool still favors TabPFN in all three regimes (Figure 2a). However, seven XGB-AC

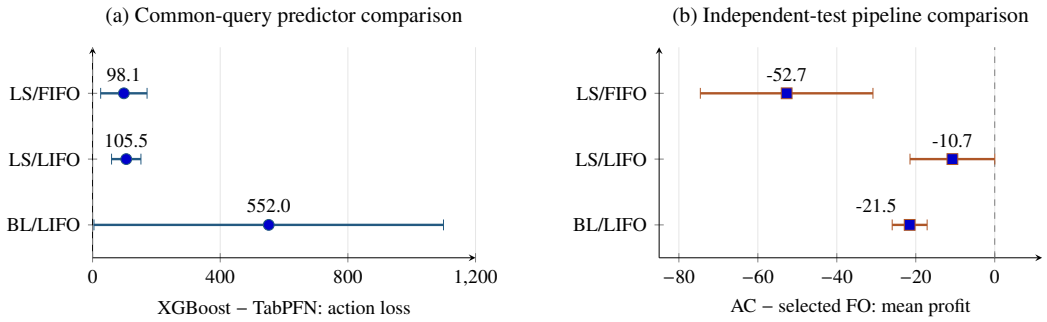

Figure 2: **A predictor advantage does not imply a pipeline advantage.** (a) XGBoost minus TabPFN one-period action-selection loss on identical histories and queries, equally averaging three source policies within each seed. (b) Decision-TabPFN minus independently selected TabPFN-FO mean profit over all 10,000 periods. Bars are paired 95% intervals across ten seeds. The panels use different metrics and horizontal scales. Source-paired contrasts in Table 1 delimit the pooled claim; the FO selection protocol is in Appendix C.

Table 1: Predictor gains depend on the query source. Full-context XGBoost minus TabPFN action-selection loss; positive values favor TabPFN. Each cell is a mean and exploratory paired 95% interval across ten seeds, after averaging 20 queries per source and seed.

Query source	LS/FIFO	LS/LIFO	BL/LIFO
AC	128.1 [79.9, 176.3]	134.2 [98.3, 170.1]	105.6 [71.2, 140.1]
FO	111.7 [55.2, 168.2]	54.0 [−48.0, 155.9]	−83.2 [−299.0, 132.6]
XGB	54.4 [−162.2, 271.0]	128.5 [−58.2, 315.2]	1633.6 [12.3, 3254.9]

backlog runs repeatedly order zero, producing extreme states that strongly affect that pool. A post hoc equal-weight AC/FO-only sensitivity leaves the BL/LIFO difference at 11.2 $[-102.6, 125.0]$. Appendix D.5 reports the source and pooled summaries; all source–context records remain in the data. The supported predictor advantage is local to the specified histories.

5.2 Does the same backbone give the same policy ranking?

Both FO candidates use the same completed observations and recover demand from operational feedback. Calibrated FO adds a validation-selected safety buffer; structured FO applies Eq. (7). Selection among six safety coefficients and structured FO chooses calibrated FO with $k = 1.4$ for LS/FIFO, and structured FO for both LIFO regimes. Test results do not change these choices.

Supplying an ordering interface changes the outcome. Selected FO exceeds AC in LS/FIFO and BL/LIFO after Holm correction (Figure 2b; Table 11 in the appendix); LS/LIFO remains unresolved at $p = 0.05034$. Figure 3 shows the operational decomposition: increased sales explain the FIFO accounting gain, while inventory-cost savings dominate in BL/LIFO. The result compares complete pipelines, with target, cost structure, and order rule changing together.

5.3 Does better context transfer to online decisions?

We compare recent, uniformly sampled, and coverage-selected contexts, each with 800 rows. Coverage selection cycles through ten observed-action bins and retrieves the closest age-inventory/backlog states within each bin. The rule jointly changes state relevance and action balance, without phase labels.

Fixed-query gains do not determine online gains. Uniform sampling lowers greedy query loss relative to recent rows in all regimes, yet lowers online profit in LS/FIFO and BL/LIFO. Offline queries hold history fixed and choose greedily; online AC generates its own history and samples among its top ten actions. The ranking change combines these protocol differences. Under the common online action rule, coverage improves over uniform sampling in both LIFO regimes, while

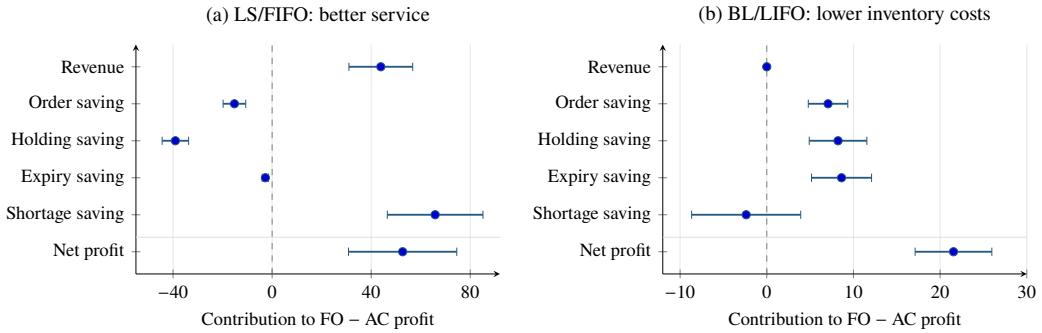

Figure 3: **Different operational trade-offs explain the two clear FO gains.** Contributions to FO minus AC mean profit over all 10,000 periods. Positive cost savings favor FO; negative savings indicate higher FO cost. Bars are paired pointwise 95% intervals across ten seeds. Component means sum to net profit, while their interval endpoints do not. The panels have different horizontal scales.

Table 2: **Fixed-query and closed-loop context rankings differ.** All contexts contain 800 rows. Left: positive recent-minus-uniform loss favors uniform sampling on common queries. Middle/right: positive profit differences favor the first online selector over 3,000 periods. Entries are exploratory paired means [pointwise 95% CI] across ten seeds. Offline queries and online policies evaluate different state distributions and action rules.

Regime	Offline loss Recent−Uniform	Online profit Uniform−Recent	Online profit Coverage−Uniform
LS/FIFO, $K = 0$	159.6 [106.4, 212.8]	−11.62 [−22.02, −1.23]	11.96 [−1.43, 25.35]
LS/LIFO, $K = 0$	148.7 [65.4, 232.0]	1.18 [−6.08, 8.44]	8.95 [0.86, 17.05]
BL/LIFO, $K = 20$	700.9 [483.3, 918.5]	−7.40 [−12.77, −2.03]	19.29 [15.11, 23.47]

FIFO is unresolved (Table 2). Phase-count-matched offline controls also favor coverage in the two LIFO regimes. They rule out phase counts alone, but do not separate state matching from action balancing (Appendix D.4).

5.4 What remains after immediate scoring is exact?

Eight finite-state instances compare exact dynamic and myopic policies with identical known demand laws, phase information, initial state, and zero terminal value. Orders are 0–8, lifetime is 2 or 3, and $T = 1{,}000$.

The dynamic policy has higher exact expected profit in every instance, including $K = 0$ (Appendix Table 22). This isolates a continuation-value gap that better immediate scores cannot remove. The separate learner–myopic gap combines prediction, information, and sampling differences; it is not a second estimate of the planning gap.

Certified short blocks leave model and boundary errors. A separate ten-seed study evaluates 1,160 tasks across four 601-action LS cells and eight small cells. With a sequence-mixture prior, checked TabPFN two-period blocks lose 347.82–410.17 profit per period to AC and 14.72–17.74 to same-prior myopic control. All twelve primary contrasts against AC, structured FO, and myopic control are negative after Holm correction. Changing only proposal ordering yields differences below 0.001, with all four intervals containing zero. In an exploratory reanalysis, sparse-prior myopic control exceeds AC by 13.87–44.50 per period across the four main cells (Appendix Table 28). Adding two-period planning then helps FIFO but hurts LIFO. Thus prior choice changes the control baseline, while the benefit of short lookahead depends on issuance. The demand mixtures use the period bound $H_{\max} = 10$; the study includes no sparse-prior TabPFN proposer.

6 From Prediction Error to Inventory Regret

The oracle comparison identifies the missing continuation value. We separate demand-prediction error from planning error using inventory structure. Write $J_T(\pi) = \mathbb{E}_\pi[\sum_{t=1}^T P_t]$ for cumulative expected profit. The guarantee concerns an explicit planning extension, not the original AC policy.

6.1 An explicit learning mechanism

Consider initially empty lost-sales inventory, zero lead time, lifetime $m \geq 1$, integer orders 0–A ($A \geq 1$), and FIFO or LIFO. Demands are independent nonnegative integers with unknown period $H \leq H_{\max}$ and finite means; revenue and cost coefficients are known and nonnegative. Demand is observed directly or recovered from shortage-cost feedback when $p > 0$. Put $\bar{d} = mA$, $d = \bar{d} + 1$, and $Z_t = \min\{D_t, \bar{d}\}$. The reward transformation

$$W_t = P_t + pD_t + c_0 \in [0, \mathcal{R}],\ c_0 = cA + K + h\bar{d} + eA, \quad \mathcal{R} = (r + p + h)\bar{d} + (c + e)A + K \tag{10}$$

changes no expected policy regret. Transitions and W_t depend only on Z_t, so unbounded demand introduces no clipping bias.

Maintain a mixture over candidate periods $1, \ldots, H_{\max}$. Each candidate uses a Dirichlet(1) categorical model for each residue and updates from every chronological demand observation. Posterior weights average these models; the true period and phase are never supplied. Write $\widehat{\nu}_t$ for this prediction and ν_t for the true clipped-demand law.

Proposition 6.1 (Prediction without period identification)**.** *The mixture satisfies*

$$\sum_{t=1}^{T} \mathbb{E}[\mathrm{KL}(\nu_t \| \widehat{\nu}_t)] \leq \mathcal{C}_T(H) \leq \log H_{\max} + H(d-1)\log(T+d). \tag{11}$$

We combine Bayesian mixture dominance and Dirichlet sequence probabilities (Hutter, 2003; Watanabe & Roos, 2015) to obtain the bound without identifying a unique period or separating phases. Appendix A.3 defines the demand-model complexity $\mathcal{C}_T(H)$. The inventory-specific step is to transfer this prediction bound without multiplying errors by the remaining horizon.

6.2 Planning with a bounded physical value effect

Let $U_t(\mathbf{v}, Z_{<t})$ be the optimal remaining value under the mixture's *joint* predictive law, with zero terminal value. Its Bellman recursion updates the mixture after each hypothetical future demand. It therefore includes future inventory costs and the value of subsequent observations.

Lemma 6.2 (Inventory value sensitivity)**.** *Put $\mathcal{G} = (r + p)mA$ and $\mathcal{S} = m\mathcal{R} + 2\mathcal{G}$. For any action-independent future demand law, including the mixture, and every remaining horizon,*

$$-m\mathcal{R} - \mathcal{G} \leq U_t(\mathbf{v}, Z_{<t}) - U_t(\mathbf{0}, Z_{<t}) \leq \mathcal{G}. \tag{12}$$

Under common orders and demands, inventory ordering is preserved and excess sales cannot exceed the initial excess stock. This gives the upper bound. Ordering zero for m periods clears any initial inventory by expiration, giving the lower bound. This includes fixed ordering costs and requires no demand-clearing event or action gap. Appendix A.11 sharpens the span to $\mathcal{S}_\sharp$, independent of K, which can replace $\mathcal{S}$ throughout; $\mathcal{R}$ still contains K.

The prediction-to-control inequality. Let $\mathbb{Q}$ be any coherent action-independent law of future clipped demand, and let $g_t^{\mathbb{Q}}$ denote the Bellman action gap along a non-anticipating policy π under that law. The proof in Appendix A.5 gives

$$J_T(\pi^\star) - J_T(\pi) \leq \underbrace{2(\mathcal{R} + \mathcal{S}) \sum_{t=1}^{T} \mathbb{E}[\mathrm{TV}(\nu_t, \widehat{\nu}_t)]}_{\text{demand-prediction error}} + \underbrace{\sum_{t=1}^{T} \mathbb{E}[g_t^{\mathbb{Q}}]}_{\text{model-planning error}}. \tag{13}$$

Here $\widehat{\nu}_t$ is the conditional prediction of $\mathbb{Q}$, and its Bellman recursion uses the complete remaining horizon. Subtract $U_{t+1}(\mathbf{0}, Z_{<t}Z_t)$ from the continuation value. The removed forecasting residual depends only on demand history, so its expectation is identical for both policies and cancels. Lemma 6.2 bounds the remaining inventory-dependent residual. This yields the coefficient $\mathcal{R} + \mathcal{S}$ instead of a coefficient proportional to the remaining horizon.

A checked decision rule. TabPFN proposes an order using the original scorer and sampler. Let $Q_t^{\mathrm{mix}}(a)$ be its immediate adjusted reward plus optimal mixture continuation. Fix a warm-up length N_0 and deterministic tolerances $\epsilon_t \geq 0$. For $t > N_0$, accept only when

$$\max_a Q_t^{\mathrm{mix}}(a) - Q_t^{\mathrm{mix}}(\widetilde{a}_t) \leq \epsilon_t; \tag{14}$$

otherwise execute a maximizing action. Denote this extension by π^{C} (Algorithm 1 in the appendix). The check concerns the learned full-horizon objective and allows any proposer; checkpoint consistency is not assumed.

6.3 A GUARANTEE FOR THE EXPLICIT EXTENSION

Theorem 6.3 (Unknown demand to optimal-policy regret)**.** *Under the preceding model, exact planning and the checked rule give, for $T \geq N_0$,*

$$J_T(\pi^\star) - J_T(\pi^{\mathrm{C}}) \leq (\mathcal{R} + \mathcal{S}) \left[N_0 + \sqrt{2T\mathcal{C}_T(H)} \right] + \sum_{t=N_0+1}^{T} \epsilon_t, \tag{15}$$

where $\pi^\star$ knows the true laws and phase and is dynamically optimal over T periods. With fixed problem parameters and $\epsilon_t = \epsilon_0/t$, expected regret is $O(\sqrt{T \log T})$.

Theorem 6.3 follows by substituting Proposition 6.1 into Eq. (13) using Pinsker's inequality. Warm-up costs at most $N_0(\mathcal{R} + \mathcal{S})$; every later model gap is at most ϵ_t. A two-point instance gives an $\Omega(\sqrt{T})$ lower bound on the time exponent, without asserting optimal dependence on other parameters.

What the bound says about implementation. For predetermined blocks, the same argument adds at most $(K_T - 1)\mathcal{S}$ at block endpoints, where K_T is the number of blocks. Growing blocks can make this charge sublinear; fixed two-period blocks cannot. Certified action gaps control only the second term of Eq. (13), not prediction error or omitted continuation. This accounts for the limits of the experiment in Section 5.4, without identifying which error dominates its observed loss. Exact count-state planning remains feasible only at small scale. All 3,040 checks passed in 160 tiny complete-horizon trajectories, with no demonstrated TabPFN-specific benefit (Appendix E.1).

Relation to exogenous-MDP guarantees. Square-root-order inventory regret is established by Gong & Simchi-Levi (2024); Wan et al. (2026) obtain rates independent of endogenous state–action cardinality in episodic Exo-MDPs. Our physical value bound holds uniformly over future demand laws in a continuing perishable system. It applies to the dependent posterior predictive law and supports cancellation under shared demand history. Known dynamics and complete-horizon planning remain assumptions. Appendix A.13 compares these mechanisms with cyclic-demand learning and inventory coupling.

7 DISCUSSION AND CONCLUSION

The predictor, ordering interface, and planner address different parts of inventory control. Same-query gains establish local scoring quality; same-backbone comparisons test the supplied decision structure; dynamic-oracle comparisons expose omitted continuation value. The inventory-value bound connects demand error and planning error without a remaining-horizon multiplier. The prior ablations further show why a valid planning certificate alone cannot select a successful demand model.

The practical implication is to compare structured demand baselines before adding a learned proposer, and to evaluate continuation error separately from action-score quality. Counterfactual labels and matched action rules remain the key tests of interface mechanisms. The guarantee covers full-feedback lost sales; scalable planning, backlog, and censored feedback require further analysis.

AI USE STATEMENT

AI helps us polish our articles and correct grammatical errors. The authors take responsibility for all content and results.

ETHICS STATEMENT

This work uses synthetic demand and simulated inventories, with no human participants or personal customer data. Deployment requires validation of service and waste outcomes beyond the profit objective.

REPRODUCIBILITY STATEMENT

Proofs and protocols appear in the appendices. Supplementary materials include code, seeds, checkpoint hashes, recorded trajectories, and reproducible mathematical checks and statistical analyses.

REFERENCES

Xiuli Chao, Xiting Gong, Cong Shi, and Huanan Zhang. Approximation algorithms for perishable inventory systems. *Operations Research*, 63(3):585–601, 2015. doi: 10.1287/opre.2015.1386.

Boxiao Chen, Jiashuo Jiang, Jiawei Zhang, and Zhengyuan Zhou. Learning to order for inventory systems with lost sales and uncertain supplies. *Management Science*, 70(12):8631–8646, 2024. doi: 10.1287/mnsc.2022.02476. URL https://doi.org/10.1287/mnsc.2022.02476.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. TabArena: A living benchmark for machine learning on tabular data. In *Advances in Neural Information Processing Systems*, volume 38, pp. 17285–17350, 2025. doi: 10.52202/085713-0519. URL https://doi.org/10.52202/085713-0519.

Joren Gijsbrechts, Robert N. Boute, Jan A. Van Mieghem, and Dennis J. Zhang. Can deep reinforcement learning improve inventory management? performance on lost sales, dual-sourcing, and multi-echelon problems. *Manufacturing & Service Operations Management*, 24(3):1349–1368, 2022. doi: 10.1287/msom.2021.1064.

Xiao-Yue Gong and David Simchi-Levi. Bandits atop reinforcement learning: Tackling online inventory models with cyclic demands. *Management Science*, 70(9):6139–6157, 2024. doi: 10.1287/mnsc.2023.4947.

Arthur Guez, David Silver, and Peter Dayan. Efficient Bayes-Adaptive reinforcement learning using sample-based search. In *Advances in Neural Information Processing Systems*, volume 25, 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/35051070e572e47d2c26c241ab88307f-Paper.pdf.

Devansh Gupta, Shiv Tavker, Dmitry Efimov, Suchitra Sathyanarayana, Gitanjali Bhutani, and Boris N. Oreshkin. Bootstrap-Conditioned Action Selection with Tabular Foundation Models, 2026. URL https://arxiv.org/abs/2608.06559. arXiv preprint.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Woonghee Tim Huh, Ganesh Janakiraman, John A. Muckstadt, and Paat Rusmevichientong. An adaptive algorithm for finding the optimal base-stock policy in lost sales inventory systems with censored demand. *Mathematics of Operations Research*, 34(2):397–416, 2009. doi: 10.1287/moor.1080.0367.

Marcus Hutter. Convergence and loss bounds for Bayesian sequence prediction. *IEEE Transactions on Information Theory*, 49(8):2061–2067, 2003. doi: 10.1109/TIT.2003.814488. URL https://www.hutter1.net/ai/spupper.htm.

Itir Z. Karaesmen, Alan Scheller-Wolf, and Borga Deniz. Managing perishable and aging inventories: Review and future research directions. In Karl G. Kempf, Pınar Keskinocak, and Reha Uzsoy (eds.), *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook, Volume 1*, volume 151 of *International Series in Operations Research & Management Science*, pp. 393–436. Springer, New York, NY, 2011. doi: 10.1007/978-1-4419-6485-4_15.

Sohom Mukherjee, Anh-Duy Pham, Richard Pibernik, and Yunbei Xu. In-Context Learning for Data-Driven Censored Inventory Control, 2026. URL https://arxiv.org/abs/2605.14840. arXiv preprint.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Steven Nahmias. Perishable inventory theory: A review. *Operations Research*, 30(4):680–708, 1982. doi: 10.1287/opre.30.4.680.

Evan L. Porteus. *Foundations of Stochastic Inventory Theory*. Stanford University Press, 2002. doi: 10.1515/9781503619883.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Herbert Scarf. The optimality of (S, s) policies in the dynamic inventory problem. In Kenneth J. Arrow, Samuel Karlin, and Patrick Suppes (eds.), *Mathematical Methods in the Social Sciences, 1959*, pp. 196–202. Stanford University Press, Stanford, CA, 1960.

Sean R. Sinclair, Felipe Vieira Frujeri, Ching-An Cheng, Luke Marshall, Hugo De Oliveira Barbalho, Jingling Li, Jennifer Neville, Ishai Menache, and Adith Swaminathan. Hindsight learning for MDPs with exogenous inputs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31877–31914, 2023. URL https://proceedings.mlr.press/v202/sinclair23a.html.

Kaustubh Sridhar, Souradeep Dutta, Dinesh Jayaraman, and Insup Lee. REGENT: A retrieval-augmented generalist agent that can act in-context in new environments. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.04759.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174, 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Jia Wan, Sean R. Sinclair, Devavrat Shah, and Martin J. Wainwright. Exploiting exogenous structure for sample-efficient reinforcement learning, 2026. URL https://arxiv.org/abs/2409.14557v4. arXiv preprint, version 4, revised July 2026.

Kazuho Watanabe and Teemu Roos. Achievability of asymptotic minimax regret by horizon-dependent and horizon-independent strategies. *Journal of Machine Learning Research*, 16(71): 2357–2375, 2015. URL https://jmlr.org/papers/v16/watanabe15a.html.

Hao Yuan, Qi Luo, and Cong Shi. Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs. *Management Science*, 67(10):6089–6115, 2021. doi: 10.1287/mnsc.2020.3799.

Huanan Zhang, Xiuli Chao, and Cong Shi. Technical note—perishable inventory systems: Convexity results for base-stock policies and learning algorithms under censored demand. *Operations Research*, 66(5):1276–1286, 2018. doi: 10.1287/opre.2018.1724.

A End-to-End Guarantee for the Planning Extension

Algorithm 1 Planning extension with checked TabPFN proposals

Require: Candidate-period bound $H_{\max}$, horizon T, warm-up length N_0, deterministic tolerances $\epsilon_t \geq 0$, original proposer

Initialize zero demand counts and uniform candidate-period weights
for $t = 1$ to T **do**
  Obtain proposal $\widetilde{a}_t$ from the original warm-up or TabPFN rule
  if $t > N_0$ **then**
    Compute Q_t^{mix} by Bellman planning with future posterior updates
    If Eq. (14) fails, replace $\widetilde{a}_t$ by a maximizing action
  end if
  Execute $\widetilde{a}_t$; observe feedback and recover $Z_t = \min\{D_t, mA\}$
  Update weights using pre-update predictions: $w_h \propto w_h q_{h,t}(Z_t)$
  Increment all candidate-period counts; append the executed transition to the context
end for

This appendix analyzes the explicit complete-horizon planning extension. The direct-scoring policies in Section 5 use a different decision rule; Algorithm 2 specifies their implementation. Complete-horizon execution at tiny scale and two-period benchmark implementations are reported separately in Appendices E.1–E.3; the latter retain the block-boundary penalty.

A.1 Information and exact bounded reduction

Assume lost sales, zero lead time, lifetime $m \geq 1$, integer orders $\{0, \ldots, A\}$ with $A \geq 1$, and fixed FIFO or LIFO issuance. Revenue and cost coefficients are known and nonnegative. Demands are independent, nonnegative integers with finite means and unknown period $H \leq H_{\max}$. The algorithm knows an upper bound $H_{\max}$, the physical transitions, and the cost coefficients, but not H, the phase laws, or their parameters. Full demand is observed after each action: under the paper's feedback model, $D_t = y_t + c_t^u/p$ when $p > 0$. Alternatively, direct demand observation suffices. Sales alone do not suffice.

Starting empty, every residual cohort is at most A. Put $\bar{d} = mA$, $d = \bar{d} + 1$, and $Z_t = \min\{D_t, \bar{d}\}$. Write $F(\mathbf{v}, a, z)$ for the deterministic age transition. Define

$$c_0 = cA + K + h\bar{d} + eA, \qquad W(\mathbf{v}, a, Z_t) = P_t + pD_t + c_0, \tag{A.1}$$

$$\mathcal{R} = (r + p + h)\bar{d} + (c + e)A + K, \qquad \mathcal{G} = (r + p)\bar{d}, \qquad \mathcal{S} = m\mathcal{R} + 2\mathcal{G}. \tag{A.2}$$

With $I = \sum_{i=2}^{m} v_i$, the adjusted reward is computable from Z_t alone:

$$W(\mathbf{v}, a, z) = (r + p)\min\{I + a, z\} - ca - K\mathbf{1}\{a > 0\} - h(I + a - z)^+ - ev_1 + c_0. \tag{A.3}$$

Thus $0 \leq W \leq \mathcal{R}$. If $D_t \geq \bar{d}$, all available stock is sold and the next state is empty, so clipping changes neither the transition nor W. This is an exact sufficient reduction for control, not an approximation to the demand tail. Since $\sum_t (pD_t + c_0)$ has the same expectation under every policy, expected regrets for P_t and W_t agree exactly. Finite demand means ensure that the original expected profits exist.

A.2 A physical value bound for any predictive law

The following argument applies to any action-independent law of future clipped demands, including a posterior predictive law whose successive draws are dependent. Let $U_t(\mathbf{v}, z_{<t})$ be its optimal remaining expected adjusted reward, with zero terminal value.

Proof of Lemma 6.2. First compare a system starting from $\mathbf{v} \geq \mathbf{0}$ with an empty system. Use identical demands and identical orders. The age transition preserves componentwise order under both

issuance rules. For example, a FIFO residual in cohort i after service is

$$\min\left\{w_i, \left(\sum_{j\le i} w_j - z\right)^+\right\}; \tag{A.4}$$

for LIFO the sum is over $j \ge i$. Both expressions are nondecreasing in every input coordinate. Expiration and the insertion of a common new order also preserve order.

Let E_t be total excess stock in the initially larger system. Common orders cancel, giving the pathwise conservation identity

$$E_t = E_{t+1} + (y_t^+ - y_t^0) + (e_t^+ - e_t^0), \tag{A.5}$$

where both differences on the right are nonnegative. Summed additional sales are therefore at most initial excess stock, which is at most $\bar{d}$. Identical orders incur identical unit and setup costs; excess holding and expiration costs are nonnegative. Hence the larger system's cumulative adjusted-reward advantage is at most $(r+p)\bar{d} = \mathcal{G}$.

Apply this comparison to a policy optimal in the larger system. The empty system can implement its orders by maintaining a virtual larger state using observed demands. Those orders remain feasible because the action set is the same at every state. The future demand law is unchanged by this construction. Taking expectations and optimizing in the empty system proves

$$U_t(\mathbf{v}, z_{<t}) - U_t(\mathbf{0}, z_{<t}) \le \mathcal{G}. \tag{A.6}$$

For the other direction, the system starting at $\mathbf{v}$ may order zero for m periods. All its initial inventory then expires, and its adjusted rewards during the wait are nonnegative. Starting empty, any policy earns at most $m\mathcal{R}$ over the same prefix. Its terminal inventory contributes at most $\mathcal{G}$ above the empty-state continuation by Eq. (A.6). Crucially, both policies induce the same distribution of the demand history at the end of the prefix. Comparing their continuations therefore gives

$$U_t(\mathbf{0}, z_{<t}) - U_t(\mathbf{v}, z_{<t}) \le m\mathcal{R} + \mathcal{G}. \tag{A.7}$$

If fewer than m periods remain, the bound follows directly from $0 \le W \le \mathcal{R}$. Combining the two inequalities proves the lemma. No clearing event, mixing condition, or action gap is needed. The argument bounds optimal values, not the profit difference between arbitrary policies that keep using identical orders forever. □

A.3 Learning unknown cyclic demand

For each candidate period $h \in \{1, \dots, H_{\max}\}$, place independent $\mathrm{Dirichlet}(1, \dots, 1)$ priors on the d-category distribution of each residue. Assign prior mass $1/H_{\max}$ to each candidate. The phase assigned by candidate h is $1 + ((t-1) \bmod h)$; no true phase label is supplied. Let $N_{h,j,z}(t)$ count clipped observations z before t in residue j under candidate h, and $N_{h,j}(t) = \sum_z N_{h,j,z}(t)$. Its next prediction is

$$q_{h,t}(z) = \frac{N_{h,j,z}(t) + 1}{N_{h,j}(t) + d}, \qquad j = 1 + ((t-1) \bmod h). \tag{A.8}$$

With $Q_h(z_{<t})$ its integrated sequence likelihood, update the mixture weights and prediction by

$$w_{h,t} = \frac{Q_h(z_{<t})}{\sum_{h'} Q_{h'}(z_{<t})}, \qquad \widehat{\nu}_t(z \mid z_{<t}) = \sum_{h=1}^{H_{\max}} w_{h,t} q_{h,t}(z). \tag{A.9}$$

Every observed demand is included, regardless of which action was proposed or executed. A shifted time origin is absorbed by the candidate's freely estimated residue laws. The construction controls predictions without asserting recovery of a unique fundamental period.

Proof of Proposition 6.1. For a residue containing n observations with category counts $n_0, \dots, n_{d-1}$, integration under the uniform Dirichlet prior gives

$$Q_j(z_{1:n}) = \frac{(d-1)!\prod_z n_z!}{(n+d-1)!} = \binom{n+d-1}{d-1}^{-1} \binom{n}{n_0, \dots, n_{d-1}}^{-1}. \tag{A.10}$$

The true probability of this count vector is at most one. Thus the probability of any one sequence with those counts under any categorical law is at most the reciprocal of the multinomial coefficient. For the true period H, multiplication over residues yields

$$\log \frac{\mathbb{P}(Z_{1:T})}{Q_H(Z_{1:T})} \le \sum_{j=1}^{H} \log \binom{n_j(T)+d-1}{d-1}. \tag{A.11}$$

Since $\mathbb{Q} \ge Q_H / H_{\max}$ pointwise, the full mixture satisfies the pathwise sequence bound

$$\begin{aligned} \log \frac{\mathbb{P}(Z_{1:T})}{\mathbb{Q}(Z_{1:T})} \le \mathcal{C}_T(H) &:= \log H_{\max} + \sum_{j=1}^{H} \log \binom{n_j(T)+d-1}{d-1} \\ &\le \log H_{\max} + H(d-1)\log(T+d). \end{aligned} \tag{A.12}$$

Here $n_j(T)$ counts times, not visits to a policy-dependent state. Taking expectations and applying the KL chain rule gives the stated cumulative prediction bound. With total variation defined as half the ℓ_1 distance, Pinsker's inequality and Cauchy–Schwarz imply

$$\sum_{t=1}^{T} \mathbb{E}[\mathrm{TV}(\nu_t, \widehat{\nu}_t)] \le \sqrt{\frac{T\mathcal{C}_T(H)}{2}}. \tag{A.13}$$

The proof combines mixture dominance and the KL chain rule (Hutter, 2003) with Dirichlet sequence probabilities (Watanabe & Roos, 2015); the period-wise combination above is specific to our predictor. Their application here is valid despite adaptive inventory decisions because the complete demand sequence is observed and unaffected by those decisions. No approximation property of the frozen checkpoint is assumed. □

A.4 The planning and action-checking algorithm

Let $U_{T+1} = 0$. Under the joint predictive law $\mathbb{Q}$, compute

$$Q_t^{\mathrm{mix}}(\mathbf{v}, z_{<t}, a) = \sum_{z=0}^{\bar{d}} \widehat{\nu}_t(z \mid z_{<t}) \left[W(\mathbf{v}, a, z) + U_{t+1}(F(\mathbf{v}, a, z), z_{<t} z) \right], \tag{A.14}$$

$$U_t(\mathbf{v}, z_{<t}) = \max_{a \in \mathcal{A}} Q_t^{\mathrm{mix}}(\mathbf{v}, z_{<t}, a). \tag{A.15}$$

Every hypothetical next observation updates all candidate-period counts and their posterior weights before the next prediction. Replacing this law by independent samples from a frozen point forecast would define a different planner.

The original TabPFN scorer and sampling rule may propose $\widetilde{a}_t$. Fix a warm-up length N_0 and deterministic tolerances $\epsilon_t \ge 0$. For $t > N_0$, accept this proposal only if

$$U_t(\mathbf{v}_{t-1}, Z_{<t}) - Q_t^{\mathrm{mix}}(\mathbf{v}_{t-1}, Z_{<t}, \widetilde{a}_t) \le \epsilon_t; \tag{A.16}$$

otherwise execute a maximizing action in Eq. (A.14). This permits the positive-temperature proposer, but does not permit an unchecked random action after certification. The proposal may depend on the complete operational history; its statistical accuracy is not a premise of the theorem. Any feasible warm-up orders are allowed, with their cost explicitly included; the main setting retains its four original warm-up orders.

The reference implementation enumerates the finite decision tree, including posterior updates, and supports rational arithmetic for small verification cases. A finite tree makes the rule constructive; it does not make it computationally efficient. Inexact planning is admissible only with valid bounds $\underline{Q}_t(a) \le Q_t^{\mathrm{mix}}(a) \le \overline{Q}_t(a)$: the check

$$\max_a \overline{Q}_t(a) - \underline{Q}_t(\widetilde{a}_t) \le \epsilon_t \tag{A.17}$$

then suffices. If it fails, refinement must continue or the actual certified gap must be charged. An arbitrary finite-depth rollout with an unverified value model does not inherit the theorem.

A.5 Cancellation of the common forecasting error

The pivotal step is to center continuation values at empty inventory. Denote the true clipped-demand law at time t by ν_t. Define the nonnegative model action gap

$$A_t^{\mathrm{mix}}(\mathbf{v}, z_{<t}, a) = U_t(\mathbf{v}, z_{<t}) - Q_t^{\mathrm{mix}}(\mathbf{v}, z_{<t}, a). \tag{A.18}$$

The Bellman residual when the model value is evaluated under the true law is

$$\begin{aligned}\delta_t(\mathbf{v}, a, z_{<t}) &= (\nu_t - \widehat{\nu}_t)\,[W(\mathbf{v}, a, Z) + U_{t+1}(F(\mathbf{v}, a, Z), z_{<t}Z)] \\ &= c_t(z_{<t}) + b_t(\mathbf{v}, a, z_{<t}),\end{aligned} \tag{A.19}$$

where the notation $(\nu - \widehat{\nu})[f]$ means the difference of expectations,

$$\begin{aligned}c_t(z_{<t}) &= (\nu_t - \widehat{\nu}_t)[U_{t+1}(\mathbf{0}, z_{<t}Z)], \\ b_t(\mathbf{v}, a, z_{<t}) &= (\nu_t - \widehat{\nu}_t)[W(\mathbf{v}, a, Z) + U_{t+1}(F(\mathbf{v}, a, Z), z_{<t}Z) - U_{t+1}(\mathbf{0}, z_{<t}Z)].\end{aligned} \tag{A.20}$$

Although c_t may be large, it depends only on the exogenous history and has the same expectation under all policies. By Lemma 6.2, the integrand defining b_t lies in $[-m\mathcal{R} - \mathcal{G}, \mathcal{R} + \mathcal{G}]$, an interval of width $\mathcal{R} + \mathcal{S}$. Consequently,

$$|b_t(\mathbf{v}, a, z_{<t})| \le (\mathcal{R} + \mathcal{S})\,\mathrm{TV}(\nu_t, \widehat{\nu}_t). \tag{A.21}$$

The bound is uniform over states and actions, including those selected from adaptive contexts. Bounding the uncentered continuation by $T\mathcal{R}$ would lose this horizon-independent control.

Proof of Theorem 6.3. Write $J_T^W(\sigma) = \mathbb{E}_\sigma[\sum_{t=1}^T W_t]$. For any non-anticipating policy σ, telescope U_t along its trajectory under the true demand law:

$$J_T^W(\sigma) - U_1(\mathbf{0}, \emptyset) = \sum_t \mathbb{E}_\sigma[-A_t^{\mathrm{mix}} + c_t(Z_{<t}) + b_t]. \tag{A.22}$$

Apply this identity to the informed optimal policy $\pi^\star$ and the checked policy π^{C}. Their demand histories have exactly the same distribution, so the expected sums of c_t cancel. The oracle's model action gaps are nonnegative, and each post-warm-up gap of π^{C} is at most ϵ_t. At a warm-up decision, any action gap is at most $\mathcal{R} + \mathcal{S}$: for each possible demand, immediate rewards differ by at most $\mathcal{R}$, and continuation values at the common updated history differ by at most $\mathcal{S}$. Thus

$$J_T^W(\pi^\star) - J_T^W(\pi^{\mathrm{C}}) \le N_0(\mathcal{R} + \mathcal{S}) + \sum_{t=N_0+1}^{T} \epsilon_t + 2(\mathcal{R} + \mathcal{S}) \sum_t \mathbb{E}[\mathrm{TV}(\nu_t, \widehat{\nu}_t)]. \tag{A.23}$$

Equation (A.13) proves the stated bound. Expected regret is identical for original and adjusted rewards. For $T < N_0$, replace N_0 by T. No planning-gap term is dropped: the comparator throughout this proof is the globally optimal finite-horizon dynamic policy, knowing the laws and phase but not future demand realizations. □

A.6 Anytime blocks and computational scope

Write $R_T(\pi) = J_T(\pi^\star) - J_T(\pi)$. For predetermined blocks, use the same globally updated mixture but plan to the end of the current block, with zero terminal value. Within each block, the preceding cancellation proof compares with its true-law optimal policy from the learner's actual starting state. Comparing this block optimum with the global optimum costs at most $\mathcal{S}$ at each nonfinal boundary, since the omitted true optimal continuation has physical value span at most $\mathcal{S}$. Telescoping across blocks gives

$$R_T(\pi^{\mathrm{C}}) \le (\mathcal{R} + \mathcal{S}) \left[N_0 + \sqrt{2T\mathcal{C}_T(H)}\right] + \sum_{t=N_0+1}^{T} \epsilon_t + (K_T - 1)\mathcal{S}, \tag{A.24}$$

when T is a block endpoint and K_T is the number of blocks. At a time inside a planned block, comparing the truncated prefix adds at most one further $\mathcal{S}$: the continuation at the truncation has that state span. Equivalently, bound the prefix action gaps using the same model potential and retain its terminal state-dependent value; the history-only terminal baseline cancels. Block lengths $1, 2, 4, \ldots$

give $K_T = O(\log T)$ and an anytime $\widetilde{O}(\sqrt{T})$ expected-regret guarantee for fixed inventory and period-class parameters. Fixed short blocks instead retain a potentially linear boundary term and are not an asymptotically optimal substitute.

The statistical result does not imply a practical planning runtime. There are up to $(A+1)^m$ physical states, and a direct belief tree has d^ℓ demand prefixes over a block of length ℓ. Appendix A.12 bounds the number of sufficient-count states, but its dimension-dependent exponent is prohibitive at the benchmark scale. In particular, the 601-action, 10,000-period benchmark has not been run with this exact planner. The proposal check can use any scorer, so neither the regret theorem nor its numerical verification establishes a TabPFN-specific advantage.

A.7 Finite budgets, failed checks, and sequence-prior robustness

The exact-planning theorem separates statistical learning from computation. The following accounting makes this separation operational: a budget stop need not invalidate a run, but its remaining action gap must be retained. No prediction accuracy of the proposal network enters this argument.

Suppose a planner for the current predetermined block maintains simultaneous bounds $L_t(a) \le Q_t^{\mathbb{Q}}(a) \le U_t(a)$ for *every* feasible action. Bounds for actions not yet examined are required. If it executes a_t, define

$$\Gamma_t = \min\Big\{\mathcal{R}+\mathcal{S}, \max\Big\{0, \max_a U_t(a) - L_t(a_t)\Big\}\Big\}. \tag{A.25}$$

The cap uses the physical action-value span, rather than an empirical estimate. At each common next demand, the immediate-reward difference is at most $\mathcal{R}$ and the continuation difference is at most $\mathcal{S}$, so every true model action gap is at most $\mathcal{R}+\mathcal{S}$.

Corollary A.1 (Budgeted planning with an explicit failure charge)**.** *Let the demand sequence model be $\mathbb{Q} = w_0\mathbb{Q}_{\text{safe}} + (1-w_0)\mathbb{Q}_{\text{alt}}$, where $0 < w_0 \le 1$, $\mathbb{Q}_{\text{safe}}$ is the period mixture in Proposition 6.1, and $\mathbb{Q}_{\text{alt}}$ is any proper action-independent sequence law. Assume all the planner's action bounds at time t hold simultaneously except on an event of conditional probability at most δ_t. At a block endpoint,*

$$\begin{aligned} R_T \le (\mathcal{R}+\mathcal{S})\left[N_0 + \sqrt{2T\{C_T(H) + \log(1/w_0)\}}\right] + (K_T - 1)\mathcal{S} \\ + \sum_{t=N_0+1}^{T} \mathbb{E}[\Gamma_t] + (\mathcal{R}+\mathcal{S}) \sum_{t=N_0+1}^{T} \mathbb{E}[\delta_t]. \end{aligned} \tag{A.26}$$

Here the reported gap is retained also at a timeout. Deterministic valid intervals have $\delta_t = 0$. Inside a planned block, add at most one further $\mathcal{S}$ as in Appendix A.6.

Proof. On the simultaneous-coverage event, the model action gap is no larger than Γ_t. On its complement it is no larger than $\mathcal{R}+\mathcal{S}$. Conditioning on the decision history therefore bounds its expectation by $\mathbb{E}[\Gamma_t] + (\mathcal{R}+\mathcal{S})\mathbb{E}[\delta_t]$. Substitute this into the centered telescoping proof and the block comparison. Finally, the pointwise sequence inequality $\mathbb{Q}(z_{1:T}) \ge w_0\mathbb{Q}_{\text{safe}}(z_{1:T})$ increases the sequence log-loss bound by at most $\log(1/w_0)$. The KL chain rule and Pinsker's inequality then give Eq. (A.26). □

For estimated values with a valid uniform error bound $|\widetilde{Q}_t(a) - Q_t^{\mathbb{Q}}(a)| \le e_t$, selecting an action within η_t of the largest estimate gives a gap at most $\eta_t + 2e_t$. This is useful only when e_t is justified simultaneously over actions and over adaptive stopping; a rollout depth, a Monte Carlo standard error for one chosen action, or an uncalibrated network score does not provide that guarantee. A confidence parameter must cover the complete adaptive search, not merely one fixed query. On a budget stop, choosing the action with the greatest lower bound minimizes the available gap certificate. A proposal can instead be accepted as soon as its own certificate passes; unchecked randomization afterwards is inadmissible.

Equation (A.26) is an accounting result, not a claim that a fixed computing budget makes the computational terms small. A sublinear conclusion additionally requires sublinear cumulative gaps and failure charges, and a sublinear number of block boundaries. For fixed block length, the displayed boundary charge may be linear even when every within-block computation is exact.

Reducing prior mass without discarding the safe component. The uniform Dirichlet prior contributes d pseudo-observations to each candidate residue. An alternative uses $\mathrm{Dirichlet}(1/d, \ldots, 1/d)$, with total mass one. We compare these individually and their equal-weight *sequence* mixture. In the last case, the candidate index is the pair consisting of period and prior type; its posterior weight is updated by the likelihood of every observation, including hypothetical observations in the planning tree. The corollary applies with $w_0 = 1/2$ and an additional complexity of $\log 2$. It does not assert that the low-mass prior always predicts better. Nor does it justify blending the two one-step forecasts with a fixed weight at every date: that operation defines a different sequence law and can accumulate a different complexity penalty.

A.8 An exact two-period block calculation

A small fixed block provides a tractable implementation for measuring proposal cost and action certificates, while retaining its boundary penalty. We give the two-period calculation explicitly so that a short-block experiment cannot be mistaken for solving the full-horizon Bellman problem.

For a next-demand distribution q, define, for total available stock s,

$$f_q(s) = (r+p)s - (r+p+h)\mathbb{E}_q[(s-Z)^+] + c_0. \tag{A.27}$$

At the last step, an inventory with oldest cohort v_1 and usable stock $I = \sum_{i=2}^{m} v_i$ has optimal value

$$\begin{aligned} V_q(\mathbf{v}) &= M_q(I) - ev_1, \\ M_q(I) &= \max\left\{ f_q(I),\ cI - K + \max_{s\in\{I+1,\ldots,I+A\}} [f_q(s) - cs] \right\}. \end{aligned} \tag{A.28}$$

The separate zero-order term is necessary when $K > 0$. Two cumulative sums compute all expected excess-stock values, and a sliding maximum computes all positive-order terms in Eq. (A.28) in $O(d)$ arithmetic operations for one distribution.

For each possible first demand z, update the posterior to obtain q_z, the law of the second demand. Compute the envelope M_{q_z} once, independently of the root order. If $\mathbf{v}'(a,z) = F(\mathbf{v}, a, z)$, then the first-step values are

$$Q^{(2)}(a) = \sum_{z=0}^{\bar{d}} \widehat{\nu}_t(z) \left[W(\mathbf{v}, a, z) + M_{q_z}\left(\sum_{i=2}^{m} v_i'(a,z) \right) - ev_1'(a,z) \right]. \tag{A.29}$$

For FIFO, $v_1' = (v_2 - z)^+$; for LIFO, $v_1' = \min\{v_2, (I + a - z)^+\}$, when $m \geq 2$. These formulas and total residual stock determine the second-step envelope index without enumerating physical states.

With C mixture components, constructing the conditional prediction matrix costs $O(Cd^2)$ by direct mixture arithmetic, the envelopes cost $O(d^2)$, and evaluating every root action costs $O(Ad)$. The conditional matrix and envelopes require $O(d^2)$ working storage, in addition to period counts. This polynomial calculation applies to a block of length two; it is not a polynomial-time algorithm for growing block lengths.

Certificates and the role of a proposal. Outward enclosures of the probabilities and arithmetic in Eqs. (A.28)–(A.29) provide action intervals. Before evaluating a root action, its expected immediate adjusted reward is a lower bound, because future adjusted rewards are nonnegative. Adding the expected maximum last-step envelope over feasible usable stock, with expiration cost omitted, supplies an upper bound. Thus all root actions have bounds even before search. A proposal orders refinement; it neither replaces these bounds nor changes the posterior. The solver checks its gap and otherwise refines or returns the actual budget-limited gap. The total decision time includes proposal inference, common envelope construction, and refinement. If common work dominates, or all root actions require refinement, a learned proposal may confer no speed advantage; this must be measured rather than inferred from the regret theorem.

A.9 A lower bound on the time exponent

An elementary two-point argument shows that the square-root time dependence cannot generally be improved. Set $m = A = 1$, $r = p = c = 1$, and $K = h = e = 0$, with fully observed

independent Bernoulli demands of unknown mean $1/2 + \theta\alpha$, where $\theta \in \{-1, 1\}$ and $\alpha = 1/(8\sqrt{T})$. The informed optimal action is one under $\theta = 1$ and zero under $\theta = -1$; a wrong action loses 2α in expectation. All actions reveal the same demand sample.

For $\alpha \le 1/4$, the one-observation KL divergence between these two Bernoulli laws is at most $16\alpha^2$. Thus, before any decision $t \le T$, the total variation between the two possible demand-history laws is at most $\sqrt{8T\alpha^2} < 1/2$. A learner's state, actions, and private randomization provide no additional information beyond this history; data processing gives the same bound for any randomized decision rule. Consequently,

$$\Pr_{+}(a_t = 0) + \Pr_{-}(a_t = 1) \ge \tfrac{1}{2}, \qquad \max\{R_T^+, R_T^-\} \ge \tfrac{1}{2}\alpha T = \frac{\sqrt{T}}{16}. \tag{A.30}$$

This lower bound concerns the time exponent for the full-feedback model. It does not establish optimal dependence on lifetime, action range, costs, or period complexity, or any computational lower bound.

A.10 Verification and limits of the result

The accompanying deterministic script checks 28,404 transition, clipping, conservation, and flushing statements across lifetimes 1–3, FIFO/LIFO, and zero or positive setup costs. Twenty small models check the mixture likelihood bound, KL chain rule, optimal-value inequalities, action certificates, common-baseline cancellation, and the regret decomposition. Two use rational arithmetic for the Bellman identities. Bounded-support cases deliberately lack a demand large enough to clear every admissible stock. The largest floating-point telescoping residual is below 5×10^{-14}. In the known-demand batching counterexample, the checked planner orders $2, 0, 2, 0, \ldots$ and earns 48 over eight periods, versus 32 for immediate maximization. This verifies the continuation mechanism in that example; it is not an unknown-demand performance experiment.

These finite checks can reveal implementation and algebraic errors, but do not prove the general theorem. No checkpoint was trained or evaluated in these checks. The theory covers full-feedback lost sales with known dynamics, a true period in the candidate class, and exact or validly certified planning. Unbounded backlog, censored sales-only feedback, action-dependent demand, and unchecked approximate planning need additional analysis. The main numerical constants are conservative, so the asymptotic rate is not a tight certificate for the empirical 10,000-period profits.

A.11 A sharper inventory span from stock conservation

The earlier bound suffices for the regret rate but counts several sources of potential sales separately. Combining prefix sales with terminal stock gives a smaller constant and removes setup costs from the inventory-value span. Define

$$\begin{aligned} \mathcal{G}_\sharp &= (r+p)(m-1)A, \\ \mathcal{L}_\sharp &= (m-1)A(r+p-c)^+ + \frac{hAm(m-1)}{2} + emA, \qquad \mathcal{S}_\sharp = \mathcal{G}_\sharp + \mathcal{L}_\sharp. \end{aligned} \tag{A.31}$$

Lemma A.2 (Stock-conservation refinement)**.** *Under the physical assumptions of Lemma 6.2, for any action-independent future demand law and any remaining horizon,*

$$-\mathcal{L}_\sharp \le U_t(\mathbf{v}, z_{<t}) - U_t(\mathbf{0}, z_{<t}) \le \mathcal{G}_\sharp. \tag{A.32}$$

Consequently, every occurrence of $\mathcal{S}$ in the regret, finite-budget, and block-boundary bounds may be replaced by $\mathcal{S}_\sharp$.

Proof. Write $w = r+p$. With at least one period remaining, the oldest initial cohort expires before the first sale. The common-order coupling in Appendix A.2 therefore yields the stronger, state-dependent upper bound

$$U_t(\mathbf{v}, z_{<t}) - U_t(\mathbf{0}, z_{<t}) \le w \sum_{i=2}^{m} v_i - e v_1 \le \mathcal{G}_\sharp. \tag{A.33}$$

At zero remaining horizon both values are zero. In particular, the weaker upper bound $w \sum_i v_i$ applies also at that terminal horizon.

For the reverse comparison, let k be the smaller of $m-1$ and the remaining horizon. Starting from $\mathbf{v}$, order zero for k periods. If a continuation remains, every surviving initial unit is now in the oldest cohort and expires before the next sale. From that point the value equals the empty-state value minus the cost of this last expiration. Total holding charges on all initial units are at most $h \sum_i (i-1)v_i$; all expiration charges, including that possible continuation correction, are at most $e \sum_i v_i$. Ignoring nonnegative sales revenues gives a feasible value at least the common expected empty-state continuation, plus the kc_0 reward offsets, minus these two charges.

Compare this policy with an optimal policy starting empty. Over the same prefix, denote its total orders, sales, and expirations by O, Y, E, and its terminal stock by $\mathbf{u}$. Conservation gives

$$Y + \sum_i u_i + E = O, \qquad O \le kA. \tag{A.34}$$

Its continuation advantage over empty stock is at most $w \sum_i u_i$. Combining that advantage with its prefix adjusted rewards, and subtracting the common kc_0 offsets, gives at most

$$wY - cO + w \sum_i u_i \le (w-c)O \le kA(w-c)^+. \tag{A.35}$$

Omitted setup, holding, and expiration charges are nonnegative. Both policies see the same exogenous-history distribution at the comparison time, so their empty-state continuation expectations cancel. Hence

$$U_t(\mathbf{0}, z_{<t}) - U_t(\mathbf{v}, z_{<t}) \le (m-1)A(w-c)^+ + h \sum_i (i-1)v_i + e \sum_i v_i \le \mathcal{L}_\sharp. \tag{A.36}$$

When $m = 1$, the prefix has length zero and the comparison is just the first expiration charge. When the horizon ends during the prefix, both terminal values are zero and the same bound holds. The centered forecasting integrand consequently has width at most $\mathcal{R} + \mathcal{S}_\sharp$, and the other uses of the value span in the original proofs are unchanged. □

For the benchmark's $m = 3$, $A = 600$, and costs, $\mathcal{G}_\sharp = 30{,}000$, $\mathcal{L}_\sharp = 35{,}400$, and $\mathcal{S}_\sharp = 65{,}400$, for both $K = 0$ and $K = 20$. These improve the constants; they do not make the generic 10,000-period bound a useful numerical profit guarantee. A fixed two-period block still contributes a potentially linear boundary term. The accompanying exact checks cover 2,592 finite models and 22,896 initial-state comparisons, including correlated demand, no demand capable of clearing every admissible stock, setup cost 1,000, and unit ordering cost exceeding $r + p$. Such checks support implementation consistency and do not substitute for the proof.

A.12 Exact dynamic programming over period-count states

The raw history tree is not the only exact representation. Replacing histories by sufficient counts is standard in Bayes-adaptive planning (Guez et al., 2012). For our specified Dirichlet period models, retain the physical inventory, the number s of observed demands, and *all* counts $N_{h,j,z}$ for $1 \le h \le H_{\max}$. Counts for the two prior concentrations are shared. The integrated likelihood of each candidate, its posterior weight, and its future predictive law are functions of these counts. Thus two histories with the same count collection have identical model action values at the same physical state. The clock supplies the current residue for every candidate, and each hypothetical observation increments the appropriate cell of every period's count table. The same statement need not hold for an arbitrary alternative sequence law in Corollary A.1.

Let $n_{h,j}(s)$ be the deterministic number of visits to residue j in the first s dates, and put

$$B_s = \prod_{h=1}^{H_{\max}} \prod_{j=1}^{h} \binom{n_{h,j}(s) + d - 1}{d-1}, \qquad M = (d-1)\frac{H_{\max}(H_{\max}+1)}{2}. \tag{A.37}$$

For each residue, its count vector is a composition of $n_{h,j}(s)$ into d nonnegative parts. Multiplying these numbers over residues bounds the possible count collections, even though collections for

overlapping candidate periods are not independent and some combinations are unreachable. There are at most $(A+1)^m B_s$ inventory–count states at date s. Backward induction, or memoized recursion that merges identical keys, therefore requires at most

$$(A+1)^m \sum_{s=0}^{T-1} B_s \le T(A+1)^m (T+d)^M \tag{A.38}$$

nonterminal state backups. Each backup considers $A+1$ actions and d demand outcomes. Predictive probabilities can be calculated from the count tables, and rational priors and costs admit exact rational arithmetic.

Equation (A.38) is a count of backups, not a bound treating arbitrary-precision arithmetic as free. This backup count is polynomial in T when the inventory and period-class parameters are fixed; it is not a practical complexity guarantee jointly in those parameters. In particular, $m = 3$, $A = 600$, and $H_{\max} = 10$ give $M = 99{,}000$. Count merging can make small full-horizon instances feasible while leaving the large benchmark out of reach. It introduces no truncation penalty: when the recursion terminates at the experiment's actual horizon, the resulting checked policy is covered by the full-horizon theorem. The fast two-period calculation in Appendix A.8 and this full-horizon calculation must therefore be reported separately.

A.13 Relation to inventory learning and exogenous-MDPs

Gong & Simchi-Levi (2024) obtain $\widetilde{O}(\sqrt{T})$ regret for nondiscarding cyclic-demand lost sales through cycle-discarding learners and a bandit layer. Their result already establishes the square-root time order for this inventory-learning setting. Our argument instead controls the value of residual age inventory under any action-independent future demand law, including a dependent posterior predictive law. Finite lifetime supplies deterministic flushing; a demand realization large enough to empty every feasible stock is unnecessary.

Chen et al. (2024) treat censored demand and uncertain supply with positive lead time. Their coupling evaluates constant-order policies near steady state, and their dynamic-optimum comparison additionally uses the near-optimality of constant ordering at sufficiently long lead times. Here, full demand feedback and known transitions permit a pathwise common-order comparison. The resulting bound concerns finite-horizon optimal values over all feasible policies, rather than steady-state values within a constant-order family. The common empty-stock baseline then removes the exogenous-history term in the prediction-to-control comparison. These are different information and computational regimes; our result retains a planning oracle.

Table 3 positions the guarantee against Wan et al. (2026). Their observed-input rate is $\widetilde{O}(L\sqrt{qN})$ over N episodes of length L, with effective dimension q; their paper uses different symbols. Our T denotes a single continuing horizon and H a demand period. The rates should therefore not be compared by matching symbol names or suppressing the inventory constants.

Table 3: Scope of the theoretical comparison. The contribution here is inventory-specific value control, not removal of known dynamics or a planning oracle.

Aspect	Exo-MDP analysis	Present inventory result
Horizon	Repeated finite episodes	One continuing trajectory
External inputs	Principally i.i.d.; observed or latent	Independent cyclic demand; observed; unknown period in a known finite class
Structure	Known reward and transition maps	Known perishable dynamics and costs
Error transfer	Episode-dependent analysis	Inventory value span independent of remaining horizon; shared-history cancellation
Computation	Exact optimization of the estimated episodic model	Complete-horizon Bellman values or valid gap bounds; benchmark scalability unresolved

B Inventory Dynamics and Experimental Protocols

B.1 Event sequence and accounting

The simulator stores m residual cohorts at the end of a period, in oldest-to-youngest order. From residual inventory $\mathbf{v}_{t-1}$ and backlog b_{t-1}, period t begins with expiration $e_t = v_{t-1,1}$ and available cohorts

$$\mathbf{w}_t = (v_{t-1,2}, \ldots, v_{t-1,m}, a_t), \qquad d_t^{\mathrm{req}} = D_t + \mathbf{1}\{\text{backlog}\} b_{t-1}. \tag{A.39}$$

Allocate d_t^{req} to the entries of $\mathbf{w}_t$ from oldest to youngest for FIFO, or in reverse order for LIFO. If $z_{t,i}$ units are issued from cohort i, then

$$v_{t,i} = w_{t,i} - z_{t,i}, \qquad y_t = \sum_i z_{t,i}, \qquad u_t = d_t^{\mathrm{req}} - y_t. \tag{A.40}$$

Allocation never exceeds available inventory or the remaining requirement, so these variables are nonnegative. The next recorded backlog is $b_t = u_t$ in backlog systems and zero in lost-sales systems. Holding cost is charged on all post-service residual units, including the oldest cohort that will expire at the next period's opening. Thus

$$c_t^o = K\mathbf{1}\{a_t > 0\} + ca_t, \quad c_t^h = h\sum_i v_{t,i}, \quad c_t^u = pu_t, \quad c_t^e = ee_t,$$
$$C_t = c_t^o + c_t^h + c_t^u + c_t^e, \qquad R_t = ry_t, \qquad P_t = R_t - C_t. \tag{A.41}$$

The initial state is empty, and no terminal salvage value is added. An end-of-period expiration charge would shift part of the label by a period; such a shift must not be applied retrospectively to a policy trained on the recorded P_t.

B.2 Feature layout and policy configuration

Table 4 specifies the available default implementation used to interpret the policy. Warm-up counts refer to decisions, not training rows. At the start of the fifth period, the observations from periods one through four produce the rows $(\mathbf{h}_1, a_2, P_2)$, $(\mathbf{h}_2, a_3, P_3)$, and $(\mathbf{h}_3, a_4, P_4)$.

Table 4: Default implementation settings and the distinction between predictor and policy temperatures.

Setting	Value / convention
Order review interval	One period
Warm-up decisions	Four, rounded $\mathrm{Unif}[200, 400]$ orders
Top-B size	10
Policy temperature	$\max\{0.1, 2/(1+0.1j)\}$, j starts at zero
Standardization stabilizer	10^{-6}
TabPFN ensemble count	One estimator
TabPFN internal softmax temperature	0.5 (distinct from policy temperature)
Pretraining-limit override	Enabled in the available implementation
Default context	All completed transition rows, one-period feature lookback
Short-lifetime feature vector	Eq. (2); 15 query columns for $m = 3$
Long-lifetime branch ($m > 5$)	Total inventory replaces age vector; expiration fields omitted

For $m > 5$, the stored observation is $(a_t, c_t^o, \|\mathbf{v}_t\|_1, c_t^h, b_t, c_t^u, y_t, C_t, R_t, P_t)$, giving 11 columns after the candidate action is appended. The age-removal diagnostic uses $m = 3$ and removes zero-based columns 2, 3, and 4 from both context and query. Other variables can still encode age indirectly. In lost-sales systems the stored backlog entry is zero, but c_t^u/p reveals unmet demand. These facts preclude describing the policy as operating on censored sales alone or using full age vectors at every tested lifetime.

The controlled scripts disable persistent policy caching and start every run with an empty history. The original eight-regime benchmark averages all 10,000 periods. Older exploratory studies and their separate burn-in conventions are retained in the historical source archive; they are not pooled with the controlled suite.

B.3 Classical and phase-specific references

Base-stock policies order toward a target inventory position; (s, S) policies order toward S when the position falls below s. The position includes carried backlog when applicable. Cyclic versions use a separate target or threshold pair for each supplied phase. In the benchmark, parameters are selected using the realized demand sequence, so these policies are hindsight references. A large mean profit for such a policy neither proves its global optimality in the perishable model nor bounds the value of all feasible policies. Finite lifetime, age-dependent issuance, and bounded orders prevent importing non-perishable structural optimality claims without a separate argument.

B.4 Costs, demand, and horizons

Table 5 gives the main cost parameters. The main benchmark uses $m = 3$, $T = 10{,}000$, $\mathcal{A} = \{0, 1, \ldots, 600\}$, $H = 10$, and ten replications. The original longer-lifetime reference implementation uses $m = 20$; the controlled suite uses $m = 3$. Controlled studies are identified in Table 7; their scores are not pooled with the original benchmark.

Table 5: Main simulation costs. All costs and revenues use the same arbitrary monetary unit.

Parameter	Symbol	Value
Unit sale price	r	10
Setup cost	K	0 or 20
Unit ordering cost	c	3
Unit holding cost	h	1
Unit unmet-demand / backlog penalty	p	15 (LS), 3 (BL)
Unit expiration cost	e	4

Table 6: Demand distributions in cycle order. Continuous draws are truncated below at zero and rounded to nonnegative integers. The implementation imposes no finite upper truncation.

Phase	Family	Parameters
1	Gaussian	Mean 125, standard deviation 15
2	Uniform	Lower 20, upper 50
3	Scaled Beta	Shapes 8 and 2; multiply by 180
4	Gamma	Shape 2, scale 20
5	Lognormal	Log-mean 5.0, log-standard deviation 0.3
6	Poisson	Rate 50
7	Exponential	Scale 140
8	Negative binomial	20 successes, success probability 0.4
9	Weibull	Shape 2, scale 170
10	Pareto	Shape 3, minimum 30

Several demand laws have unbounded support. The exact lost-sales reduction in Appendix A.1 removes the policy-independent demand-tail term; it does not impose a cap on observed demand or extend the theorem to unbounded backlog. The controlled suite uses paired Student-t intervals after seed aggregation; the original benchmark uses the marginal convention stated in Appendix F.

B.5 Broader impacts

Improved ordering can reduce waste and shortages, but a profit objective reflects the specified cost coefficients rather than all social costs of availability and disposal. The present experiments do not establish deployment safety, robustness to supply disruptions, or benefits for a specific retail or healthcare operation. Those uses require domain-specific validation and appropriate operational constraints.

Algorithm 2 Decision-TabPFN with completed-transition feedback

Require: Frozen regressor q_θ, action set $\mathcal{A}$, horizon T, warm-up N_0, top-B size, temperature parameters

$\mathcal{D} \leftarrow \emptyset$; $j \leftarrow 0$
for $t = 1$ to T **do**
  if $t \le N_0$ **then**
    Draw a warm-up order a_t from the specified initialization distribution
  else
    Batch-score $\{\mathrm{Cat}(\mathbf{h}_{t-1}, a) : a \in \mathcal{A}\}$ using $q_\theta(\cdot \mid \mathcal{D})$
    Retain the top-B scores; compute p_t by Eqs. (4)–(5)
    Draw $a_t \sim p_t$; $j \leftarrow j + 1$
  end if
  Execute a_t and observe completed-period feedback $\mathbf{h}_t$
  if $t \ge 2$ **then**
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathrm{Cat}(\mathbf{h}_{t-1}, a_t), P_t)\}$
  end if
end for

C Controlled Evaluation Protocol

This suite strengthens the separation between the predictor, ordering rule, and context. It uses the accounting and features in Appendix B, but independently specifies validation seeds, common warm-up, demand-reconstructing FO targets, and the 601-action diagnostic grid. Earlier exploratory comparisons are archived separately and are not pooled with this suite.

C.1 Task allocation and independent units

Table 7: Complete controlled suite. A task is an execution unit, not an independent statistical sample. Each oracle task produces both a DP and a myopic trajectory.

Group	Tasks	Horizon	Design
Validation	120	1,000	3 regimes × 5 seeds × 8 configurations
Policy controls	150	3,000	3 regimes × 10 seeds × 5 policies
Online context	90	3,000	3 regimes × 10 seeds × 3 selectors
Common queries	90	20 queries	3 regimes × 10 seeds × 3 source policies
Longer cycle	90	3,000	3 regimes × 10 seeds × 3 policies
Finite-state study	160	1,000	8 instances × 10 seeds × (learner + oracle pair)
Long-horizon test	60	10,000	3 regimes × 10 seeds × 2 policies
Total	760		750 trajectories plus 90 diagnostic tasks

Validation seeds are 240100–240104 and test seeds are 240200–240209. Each seed determines a demand path shared across policy comparisons in a regime. Separate random streams generate demand, initial orders, action sampling, context subsets, and diagnostic demand draws. The test settings and context selector are fixed before their outcomes are inspected. The long-horizon runs preserve their exact 3,000-period prefixes, so the short and long results are not independent replications.

The 750 trajectories comprise 510 validation/core/context/period/long rollouts, 80 finite-state learner rollouts, and 160 finite-state oracle trajectories. Each of the 90 probe tasks evaluates 20 queries, six contexts, and two predictors, yielding 21,600 records. Queries, periods, and source policies sharing a seed are aggregated before uncertainty is estimated. Primary hypotheses concern the three long-horizon AC–selected-FO comparisons; their two-sided p-values receive Holm correction. The reported 95% intervals remain pointwise Student-t intervals, not simultaneous confidence intervals. Auxiliary comparisons use pointwise intervals without a claim of familywise significance.

C.2 Backbones, features, and action rules

All principal TabPFN variants use TabPFN 2.0.9 with one estimator, internal softmax temperature 0.5, random state 0, and the pretraining-limit override enabled. Weights remain frozen. Each learned decision refreshes the completed-transition context. Native AC uses the top-10 standardized softmax in Eq. (4); greedy AC replaces only that sampling rule with an argmax. The greedy XGB-AC control uses 100 trees, depth 6, learning rate 0.08, row and column subsampling 0.8, one CPU thread, and random state 0. It refits each period. This differs from the original Decision-XGB benchmark (50 trees, depth 5, learning rate 0.1, native softmax selection).

For $m = 3$, AC fits 15-column queries $[\mathbf{h}_{t-1}, a_t]$ to P_t. FO fits the same 14-entry observation $\mathbf{h}_{t-1}$ to realized new demand. Both first use the model in period five after four paired warm-up orders. The FO target is reconstructed only after observing the period:

$$D_t = y_t + c_t^u/p - \mathbf{1}\{\mathrm{BL}\}b_{t-1}. \tag{A.42}$$

This identity is checked against the simulated demand in every recorded rollout. It uses the operational feedback available to both pipelines and introduces no future-demand input. It does require observed shortage cost or an equivalent unmet-demand count; censored sales alone would not supply the target.

C.3 Forecast-to-order rules

Calibrated FO. Let $\widehat{d}_t$ be the nonnegative TabPFN mean forecast. A buffer statistic σ_t is the population standard deviation of the most recent at most 200 realized forecast errors, set to zero until ten such errors exist. Define

$$U_t = \begin{cases} \sum_{i=2}^{m} v_{t-1,i} - \mathbf{1}\{\mathrm{BL}\}b_{t-1}, & \text{FIFO}, \\ v_{t-1,m} - \mathbf{1}\{\mathrm{BL}\}b_{t-1}, & \text{LIFO}. \end{cases} \tag{A.43}$$

The order is $a_t = \mathrm{clip}_{[0,600]}(\mathrm{round}[\widehat{d}_t + k\sigma_t - U_t])$. For LIFO this is an inventory heuristic: intermediate cohorts remain physically available even though U_t discounts them. The grid $k \in \{0.2, 0.4, 0.6, 1.0, 1.4, 2.0\}$ is searched on validation only.

Structured FO. From the predicted demand distribution, evaluate 129 midpoint quantiles at probabilities $(j + 1/2)/129$, $j = 0, \ldots, 128$, rounding and clipping them below at zero. Let $\widehat{d}_{t,j}$ be these values. The physical one-period profit is

$$G(\mathbf{v}, b, a, d) = r\min\{S, d + b\} - ca - K\mathbf{1}\{a > 0\} - h(S - d - b)^+ - p(d + b - S)^+ - ev_1,$$
$$S = \sum_{i=2}^{m} v_i + a. \tag{A.44}$$

For lost sales, $b = 0$; for backlog, b is the outstanding requirement. Structured FO chooses

$$a_t = \arg\max_{a \in \{0,\ldots,600\}} \frac{1}{129} \sum_{j=0}^{128} G(\mathbf{v}_{t-1}, b_{t-1}, a, \widehat{d}_{t,j}). \tag{A.45}$$

The profit at a fixed pre-decision state depends on total usable stock; FIFO/LIFO changes the future age state through issuance. This optimization inserts the known inventory and cost structure into the policy. It approximates a one-period expectation, not a dynamic-programming value. A preflight comparison of 129 and 513 quantiles on four queries gave action changes of 1, 1, 0, and 0 units and expected-profit loss at most 0.0468; this is a numerical check, not a global quadrature-error bound.

The selected safety coefficient is 1.4, 0.4, and 0.2 in LS/FIFO, LS/LIFO, and BL/LIFO. Comparing each calibrated candidate against structured FO on the same validation set selects the former only for FIFO. No test mean is used to revise this choice.

C.4 Equal-budget context selection

All non-full selectors retain at most $W = 800$ preceding completed rows. Recent selection takes the last W; uniform selection samples without replacement and restores chronological order. Coverage

selection partitions observed actions into ten equal-width bins. Within each bin, rows are sorted by the mean squared distance between their age-inventory/backlog entries and the query's entries, after dividing each coordinate by 600. The algorithm cycles through the bins, taking the next available nearest row until W rows are selected. Empty bins are skipped and ties preserve original row order. The selected rows are returned in chronological order. If fewer than W rows exist, all are used.

The selector therefore balances *observed* actions; it cannot create evidence for unobserved actions. State distance and action balancing change together. The evaluation tests this joint rule, not the isolated causal effect of either ingredient. It requires neither the demand period nor labels identifying its phase.

Phase-count controls. Two additional offline selectors enforce exactly the recent window's count in each true label phase: 80 rows per phase at these queries. Phase-uniform samples within phase; phase-coverage applies the same coverage rule within phase. The row $[\mathbf{h}_{k-1}, a_k]$ has label phase ϕ_k. These controls use simulator phase information and are never deployed as online policies. Comparing them tests whether phase counts alone explain a fixed-query context effect.

C.5 Queries and Monte Carlo targets

Each source trajectory is generated by native AC, greedy XGB-AC, or structured FO for 3,000 periods. The zero-based query days are 2000–2009 and 2800–2809, corresponding to paper periods 2001–2010 and 2801–2810. At a query, only earlier completed transitions are used. The 601 candidate actions share 2,048 fresh draws from the true phase distribution, and every predictor/context variant shares the resulting expected-profit vector. A fixed seed/query random stream determines the draws. The reported intervals describe variability across seed-level losses; they do not separately estimate the Monte Carlo approximation error.

Queries are evaluated greedily because their purpose is to measure score-based action choice. In contrast, native online AC samples from its top ten scores and changes the subsequent inventory and training history. Offline and online context effects consequently have different estimands. All three source distributions and all seeds remain in the specified summaries, including high-backlog XGB trajectories.

C.6 Exact finite-state instances

The finite-state study uses lost sales, $m \in \{2, 3\}$, FIFO/LIFO, $K \in \{0, 20\}$, and orders 0–8, with other unit costs unchanged. Demand has support 0–6 and alternates between

$$(0.20, 0.40, 0.25, 0.10, 0.05, 0, 0), \qquad (0, 0, 0.05, 0.10, 0.25, 0.30, 0.30). \tag{A.46}$$

Each of the 9^m residual inventory vectors is enumerated. With horizon 1,000 and zero terminal value, backward induction computes the optimal policy and its expected value at the empty initial state. The phase-informed myopic policy maximizes exact one-period expected profit; backward evaluation computes its value under the same transition kernel. Their exact value difference divided by 1,000 is the planning gap in Table 22. Both policies are also simulated on the learner's ten demand paths to obtain paired trajectory comparisons. The learner's four warm-up draws are rounded from $\mathrm{Unif}[8/3, 16/3]$; the oracles act from the first period.

C.7 Execution and reproducibility

The suite ran on an A100 40GB GPU with at most ten concurrent workers, using Python 3.12, PyTorch 2.7.1, NumPy 1.26.4, scikit-learn 1.5.2, and XGBoost 3.0.5. Shared-GPU timings reflect contention and are not isolated deployment latency. The original implementations, frozen configurations, software versions, model fingerprint, seed-level records, and analysis scripts are retained in the supplementary materials. Completion checks verify each task's horizon and configuration hash. Independent replay checks verify inventory accounting and the finite-state Bellman calculations. The analysis does not discard low returns or select methods using test outcomes.

D Controlled Results and Source Sensitivity

This appendix retains the controlled comparisons and the qualifications needed to interpret them. Complete query-level records, all source–context summaries, and older descriptive tables remain in the accompanying data and archived analysis. Unless a table specifies otherwise, entries are means with pointwise Student-t 95% intervals across ten seeds. Paired contrasts are formed within seed. The primary long-horizon family additionally uses Holm correction; auxiliary intervals are descriptive and do not provide simultaneous coverage. All trajectories and query records are included in the supplementary data.

D.1 Independent validation and selection

The validation set has five seeds and 1,000 periods per run. The six safety coefficients are fixed in advance. Selection first chooses the coefficient with the largest validation mean within each regime and then compares that candidate with structured FO. The resulting choice is fixed for the 10,000-period test comparison. AC is shown for reference and is not part of FO selection.

Table 8: Validation mean profit over all 1,000 periods, five seeds. These intervals describe validation variability; they are not test estimates.

Candidate	LS/FIFO, $K = 0$	LS/LIFO, $K = 0$	BL/LIFO, $K = 20$
AC	325.9 [294.7, 357.2]	251.4 [175.8, 327.1]	431.7 [379.7, 483.6]
FO, $k = 0.2$	292.0 [266.9, 317.0]	289.1 [268.2, 309.9]	499.7 [483.1, 516.3]
FO, $k = 0.4$	344.7 [323.5, 365.9]	310.0 [295.1, 324.9]	473.6 [455.5, 491.6]
FO, $k = 0.6$	377.2 [357.2, 397.2]	306.4 [292.2, 320.6]	436.4 [417.4, 455.4]
FO, $k = 1$	409.8 [392.2, 427.4]	265.7 [244.8, 286.6]	348.5 [325.8, 371.3]
FO, $k = 1.4$	421.3 [408.6, 433.9]	193.5 [166.9, 220.1]	245.8 [218.6, 273.1]
FO, $k = 2$	408.7 [396.5, 421.0]	52.0 [22.0, 82.1]	83.4 [50.1, 116.7]
Structured FO	360.8 [343.2, 378.5]	333.4 [315.4, 351.5]	516.6 [506.3, 526.9]

D.2 Policy controls at 3,000 periods

The five policies use paired demand paths and initial orders. Greedy AC changes only the action rule of full-history AC. Greedy XGB-AC changes both the predictor and action rule relative to native AC; it is distinct from the original Decision-XGB benchmark configuration. The calibrated FO coefficients are 1.4, 0.4, and 0.2 in column order, selected on validation.

Table 9: All 3,000-period mean profits, including warm-up. No failed or low-return trajectory is removed.

Policy	LS/FIFO, $K = 0$	LS/LIFO, $K = 0$	BL/LIFO, $K = 20$
Full-history AC	357.9 [316.2, 399.7]	315.9 [297.1, 334.7]	490.6 [483.0, 498.2]
Greedy AC	377.8 [355.1, 400.4]	329.4 [316.5, 342.4]	493.6 [485.5, 501.7]
Calibrated FO	435.9 [431.6, 440.3]	329.8 [325.0, 334.7]	510.6 [506.3, 514.8]
Structured FO	374.5 [369.5, 379.5]	350.8 [346.3, 355.3]	530.6 [526.2, 535.0]
Greedy XGB-AC	60.3 [−293.9, 414.6]	−192.3 [−656.9, 272.3]	−285, 026.3 [−425, 952.9, −144, 099.7]

Zero-order trajectories. In BL/LIFO, seven of ten greedy XGB-AC seeds order zero throughout periods 5–3,000. Their terminal backlog is 267,990–275,179, and the per-unit-per-period backlog charge dominates their losses. The remaining three profits are 424.94, 431.90, and 403.77. Zero-order trajectories occur in one FIFO and two lost-sales LIFO seeds as well. Reconstructing inventory and costs confirms the reported returns; the precise internal model cause is unresolved. The resulting mixture makes a ten-seed Student-t interval a coarse uncertainty summary. It is evidence about this greedy configuration, not all XGBoost policies.

Table 10: Paired policy differences at 3,000 periods. Wide intervals for XGB reflect its heterogeneous closed-loop outcomes.

Contrast	LS/FIFO, $K = 0$	LS/LIFO, $K = 0$	BL/LIFO, $K = 20$
AC−greedy AC	-19.9 $[-70.2, 30.4]$	-13.5 $[-35.3, 8.2]$	-3.0 $[-9.1, 3.0]$
AC−calibrated FO	-78.0 $[-120.0, -36.1]$	-14.0 $[-35.1, 7.2]$	-20.0 $[-27.7, -12.2]$
AC−structured FO	-16.6 $[-55.9, 22.7]$	-34.9 $[-53.9, -16.0]$	-40.0 $[-47.9, -32.2]$
AC−greedy XGB	297.6 $[-70.5, 665.7]$	508.2 $[39.3, 977.1]$	$285,516.9$ $[144,586.8, 426,447.0]$

Table 11: **Independent selection changes the pipeline comparison.** All 10,000-period mean profits, including four initial random orders. FO is selected on five disjoint validation seeds: calibrated FO ($k = 1.4$) for LS/FIFO and structured FO for both LIFO regimes. Differences use ten paired test seeds; intervals are pointwise 95%, and p-values are Holm-adjusted across the three comparisons.

Regime	AC	FO	AC−FO [95% CI]	Holm p
LS/FIFO, $K = 0$	381.95	434.65	-52.70 $[-74.57, -30.84]$	8.08×10^{-4}
LS/LIFO, $K = 0$	340.77	351.49	-10.72 $[-21.46, 0.02]$	0.05034
BL/LIFO, $K = 20$	513.05	534.59	-21.54 $[-25.97, -17.12]$	4.78×10^{-6}

D.3 Long-horizon profits and cost accounting

The 60 long runs continue their 3,000-period prefixes with preserved action RNG state and history. Demand, actions, observations, and costs in the prefix are identical. Table 11 gives the paired primary test; Table 12 gives marginal uncertainty.

Table 12: All 10,000-period profits with marginal 95% intervals. The test uses the frozen validation choice in each regime.

Regime	AC	FO rule	FO
LS/FIFO, $K = 0$	381.95 [359.85, 404.05]	Calibrated FO	434.65 [431.29, 438.02]
LS/LIFO, $K = 0$	340.77 [327.88, 353.66]	Structured FO	351.49 [348.08, 354.90]
BL/LIFO, $K = 20$	513.05 [509.23, 516.87]	Structured FO	534.59 [533.24, 535.94]

D.4 Context selection in the online loop

The full-history row is shared with the core experiment. Each other policy uses its own preceding interactions and selects at most 800 rows. All use the same native top-10 sampling rule. The selector is fixed before the test runs, and no context budget or distance parameter is chosen from these results.

D.5 Common-query diagnostics and source stratification

Each source policy contributes 20 fixed queries per seed. For each model and context, losses are averaged across queries, then equally across the three sources within seed. The 21,600 model–context–query records are dependent observations, not 21,600 independent samples. Table 17 reports this specified pooled estimand. Table 1 gives paired full-context contrasts. Table 19 gives their marginal losses; all six contexts for each source are retained in `probe_source_summary.csv`.

Extreme source states. In BL/LIFO, median backlog at the queries is zero for native AC and structured FO sources but 181,402.5 for the XGB source; its maximum is 258,167. The pooled model contrast consequently depends on the equally weighted source mixture. This mixture is retained rather than changed after observing performance. A large pooled difference cannot be interpreted as a gain of that size on every source distribution.

Source-paired sensitivity. Table 1 recomputes paired predictor differences within each source from the retained query records. It introduces no new model fitting or trajectories. As a post hoc

Table 13: Mean daily cost accounting over all 10,000 periods. Order cost includes any setup charge. Profit equals revenue minus the four costs, up to rounding.

Regime	Policy	Revenue	Order	Holding	Expiry	Shortage	Profit
LS/FIFO, $K = 0$	AC	817.36	248.21	38.50	3.99	144.72	381.95
LS/FIFO, $K = 0$	FO	861.25	263.43	77.58	6.71	78.88	434.65
LS/LIFO, $K = 0$	AC	813.91	263.49	34.01	25.74	149.89	340.77
LS/LIFO, $K = 0$	FO	808.13	254.71	27.03	16.35	158.56	351.49
BL/LIFO, $K = 20$	AC	913.84	306.40	29.69	18.51	46.19	513.05
BL/LIFO, $K = 20$	FO	913.84	299.33	21.46	9.89	48.57	534.59

Table 14: Paired contributions to FO minus AC profit. Positive values increase FO profit. Contributions sum to the total at the mean level; interval endpoints do not add. This is an accounting decomposition of existing trajectories.

Contribution	LS/FIFO, $K = 0$	LS/LIFO, $K = 0$	BL/LIFO, $K = 20$
Revenue increase	$43.89\ [31.01, 56.77]$	$-5.78\ [-14.65, 3.10]$	$-0.00\ [-0.00, 0.00]$
Order-cost saving	$-15.23\ [-19.79, -10.66]$	$8.78\ [3.72, 13.85]$	$7.07\ [4.79, 9.35]$
Holding-cost saving	$-39.07\ [-44.42, -33.72]$	$6.98\ [3.25, 10.71]$	$8.23\ [4.90, 11.55]$
Expiration-cost saving	$-2.72\ [-3.98, -1.47]$	$9.40\ [6.12, 12.67]$	$8.63\ [5.16, 12.09]$
Shortage-cost saving	$65.84\ [46.52, 85.15]$	$-8.66\ [-21.98, 4.65]$	$-2.38\ [-8.67, 3.92]$
Total FO minus AC	$52.70\ [30.84, 74.57]$	$10.72\ [-0.02, 21.46]$	$21.54\ [17.12, 25.97]$

Table 15: Mean online profit across all 3,000 periods. Context size and composition are part of the evaluated policy.

Context	LS/FIFO, $K = 0$	LS/LIFO, $K = 0$	BL/LIFO, $K = 20$
Full-history AC	$357.91\ [316.15, 399.66]$	$315.87\ [297.09, 334.65]$	$490.59\ [482.96, 498.23]$
Recent 800	$356.28\ [332.24, 380.33]$	$304.10\ [294.05, 314.15]$	$477.12\ [471.12, 483.11]$
Uniform 800	$344.66\ [318.82, 370.51]$	$305.28\ [290.96, 319.59]$	$469.72\ [463.07, 476.36]$
Coverage 800	$356.62\ [322.04, 391.20]$	$314.23\ [297.89, 330.57]$	$489.00\ [479.56, 498.45]$

Table 16: All six paired context-policy comparisons. An interval containing zero does not establish equivalence.

Contrast	LS/FIFO, $K = 0$	LS/LIFO, $K = 0$	BL/LIFO, $K = 20$
Coverage−Uniform	$11.96\ [-1.43, 25.35]$	$8.95\ [0.86, 17.05]$	$19.29\ [15.11, 23.47]$
Coverage−Recent	$0.34\ [-15.40, 16.08]$	$10.13\ [0.75, 19.51]$	$11.89\ [5.60, 18.18]$
Coverage−Full	$-1.28\ [-24.50, 21.93]$	$-1.64\ [-7.55, 4.27]$	$-1.59\ [-5.74, 2.56]$
Full−Recent	$1.62\ [-24.75, 27.99]$	$11.77\ [1.01, 22.53]$	$13.48\ [8.94, 18.02]$
Full−Uniform	$13.24\ [-5.58, 32.06]$	$10.59\ [1.34, 19.85]$	$20.88\ [17.78, 23.98]$
Recent−Uniform	$11.62\ [1.23, 22.02]$	$-1.18\ [-8.44, 6.08]$	$7.40\ [2.03, 12.77]$

Table 17: Pooled one-period action-selection loss (lower is better). Full uses all available rows; every other context uses 800 rows. Phase-conditioned selectors are offline oracle diagnostics.

Predictor	Context	LS/FIFO, $K = 0$	LS/LIFO, $K = 0$	BL/LIFO, $K = 20$
TabPFN	Full	187.5 [139.5, 235.4]	231.2 [118.5, 343.9]	1,019.1 [389.0, 1,649.1]
TabPFN	Recent	419.2 [326.6, 511.7]	443.8 [309.5, 578.1]	2,138.8 [1,581.2, 2,696.3]
TabPFN	Uniform	259.6 [192.6, 326.6]	295.1 [228.2, 362.0]	1,437.9 [945.2, 1,930.6]
TabPFN	Coverage	235.8 [160.6, 311.0]	235.1 [148.6, 321.6]	589.5 [290.2, 888.9]
TabPFN	Phase-uniform	240.8 [162.1, 319.6]	317.7 [224.8, 410.5]	1,447.8 [941.4, 1,954.3]
TabPFN	Phase-coverage	208.6 [142.4, 274.8]	236.5 [132.5, 340.4]	559.7 [266.1, 853.3]
XGBoost	Full	285.5 [172.6, 398.4]	336.7 [188.0, 485.5]	1,571.1 [944.1, 2,198.0]
XGBoost	Recent	430.4 [333.7, 527.2]	430.8 [307.2, 554.5]	1,919.5 [1,322.0, 2,516.9]
XGBoost	Uniform	327.4 [214.0, 440.8]	381.0 [235.0, 527.0]	1,662.9 [1,017.6, 2,308.2]
XGBoost	Coverage	288.9 [180.2, 397.6]	330.2 [184.9, 475.4]	1,567.9 [937.1, 2,198.6]
XGBoost	Phase-uniform	330.0 [226.3, 433.7]	370.5 [233.5, 507.5]	1,704.2 [1,094.2, 2,314.3]
XGBoost	Phase-coverage	304.6 [178.4, 430.9]	344.2 [183.1, 505.3]	1,548.1 [902.6, 2,193.6]

Table 18: Paired pooled diagnostic contrasts. Positive model differences favor TabPFN; positive context differences favor the second context. Intervals are pointwise.

Contrast	LS/FIFO, $K = 0$	LS/LIFO, $K = 0$	BL/LIFO, $K = 20$
XGB - TabPFN (full)	98.1 [25.4, 170.8]	105.5 [59.5, 151.6]	552.0 [4.4, 1,099.6]
XGB - TabPFN (recent)	11.2 [−30.6, 53.1]	−13.0 [−60.0, 34.0]	−219.3 [−441.1, 2.4]
XGB - TabPFN (uniform)	67.8 [10.0, 125.7]	85.9 [−7.6, 179.5]	225.0 [30.9, 419.1]
XGB - TabPFN (coverage)	53.2 [−6.6, 112.9]	95.1 [30.5, 159.6]	978.3 [462.5, 1,494.2]
XGB - TabPFN (phase-uniform)	89.2 [46.8, 131.5]	52.8 [−32.5, 138.0]	256.4 [100.8, 412.0]
XGB - TabPFN (phase-coverage)	96.0 [22.1, 169.9]	107.7 [36.0, 179.5]	988.4 [468.6, 1,508.1]
recent - uniform (TabPFN)	159.6 [106.4, 212.8]	148.7 [65.4, 232.0]	700.9 [483.3, 918.5]
uniform - coverage (TabPFN)	23.8 [−15.0, 62.7]	60.0 [15.8, 104.2]	848.4 [462.1, 1,234.6]
phase-uniform - phase-coverage (TabPFN)	32.2 [−15.8, 80.3]	81.2 [12.4, 150.1]	888.1 [484.0, 1,292.3]
recent - uniform (XGB)	103.0 [60.7, 145.3]	49.8 [14.0, 85.6]	256.6 [125.9, 387.2]
uniform - coverage (XGB)	38.5 [6.4, 70.5]	50.8 [21.7, 80.0]	95.0 [59.5, 130.6]
phase-uniform - phase-coverage (XGB)	25.4 [−11.7, 62.5]	26.3 [−18.7, 71.3]	156.2 [95.7, 216.6]

Table 19: Full-context action-selection loss by source. Means and marginal 95% intervals across ten seeds; each seed averages 20 queries. Paired predictor differences appear in the main text.

Source	Predictor	LS/FIFO	LS/LIFO	BL/LIFO
AC	TabPFN	71.3 [32.5, 110.2]	63.4 [50.7, 76.0]	72.5 [40.9, 104.1]
AC	XGBoost	199.5 [149.1, 249.8]	197.6 [162.5, 232.6]	178.2 [164.2, 192.1]
FO	TabPFN	139.0 [95.2, 182.8]	198.7 [88.9, 308.6]	358.2 [113.9, 602.4]
FO	XGBoost	250.7 [222.1, 279.3]	252.7 [214.2, 291.1]	275.0 [112.5, 437.4]
XGB	TabPFN	352.0 [204.3, 499.7]	431.5 [165.1, 697.9]	2626.5 [754.8, 4498.2]
XGB	XGBoost	406.4 [65.4, 747.4]	559.9 [110.2, 1009.6]	4260.1 [2278.9, 6241.3]

sensitivity, equal weighting of only AC and FO sources gives XGBoost-minus-TabPFN losses of 119.9 $[78.2, 161.6]$, 94.1 $[40.8, 147.3]$, and 11.2 $[-102.6, 125.0]$ in regime order. The original equal-three-source estimand and all extreme trajectories remain reported. These intervals are exploratory and pointwise; source reweighting does not establish an intrinsic advantage under arbitrary deployment distributions.

D.6 A 20-PERIOD DEMAND CYCLE

Each of the same ten demand families lasts two consecutive periods. The policies restart from four warm-up orders and learn online; no policy learned at period 10 is frozen and transferred. Model and context settings are unchanged. The tested FO is the pre-specified structured rule; a calibrated FO sweep is not repeated in this stress test.

Table 20: Mean profit over all 3,000 periods under the longer demand cycle.

Policy	LS/FIFO, $K = 0$	LS/LIFO, $K = 0$	BL/LIFO, $K = 20$
Full-history AC	191.94 $[158.60, 225.29]$	115.81 $[37.32, 194.30]$	471.21 $[463.08, 479.35]$
Coverage 800	202.86 $[164.25, 241.47]$	114.95 $[33.43, 196.47]$	468.43 $[459.11, 477.76]$
Structured FO	168.77 $[156.57, 180.97]$	161.02 $[150.93, 171.11]$	498.90 $[495.63, 502.17]$

Table 21: Paired contrasts under the 20-period cycle.

Contrast	LS/FIFO, $K = 0$	LS/LIFO, $K = 0$	BL/LIFO, $K = 20$
AC−structured FO	23.17 $[-10.73, 57.08]$	-45.21 $[-129.92, 39.50]$	-27.69 $[-35.61, -19.77]$
AC−coverage	-10.92 $[-40.34, 18.51]$	0.86 $[-27.90, 29.61]$	2.78 $[-1.20, 6.76]$

D.7 FINITE-STATE ORACLE COMPARISONS

Table 22: **One-period optimal actions leave a planning gap.** AC, myopic oracle, and DP columns give all-period mean profits on ten paired demand paths. The last column is the exact expected DP-minus-myopic value divided by 1,000; it uses full state enumeration, not sampled trajectories. Both oracles know the same demand law and phase. Paired intervals appear in Appendix D.7.

m	Issue	K	AC	Myopic	DP	Exact gap
2	FIFO	0	13.61	16.46	17.46	1.105
2	FIFO	20	-0.42	2.21	2.86	0.705
2	LIFO	0	11.23	15.53	15.70	0.194
2	LIFO	20	-2.76	1.02	2.86	1.892
3	FIFO	0	16.15	18.35	19.32	1.068
3	FIFO	20	2.22	4.06	7.76	3.762
3	LIFO	0	13.11	16.97	17.38	0.456
3	LIFO	20	-0.83	2.23	5.18	3.029

Table 22 distinguishes simulated mean profits from exact expected planning gaps. Here we report paired differences over the same ten demand paths. The DP and myopic oracles have identical information; their difference isolates planning. The myopic-oracle minus learner comparison also contains information, learning, warm-up, and action-sampling effects and is not a pure prediction-error measure.

The exact per-period planning gap ranges from 0.194 to 1.105 with $K = 0$, and from 0.705 to 3.762 with $K = 20$. A positive gap without setup cost shows that one-period optimization can remain suboptimal under carryover and expiration even when ordering has no fixed charge. The nine-action setting also means that the top-10 learner retains all nine actions, which limits direct extrapolation of its return to the larger action space.

Table 23: Paired all-period profit differences on the finite-state problems. Exact expectation gaps are reported separately in Table 22.

Instance	DP minus myopic	Myopic minus AC	DP minus AC
$m = 2$, FIFO, $K = 0$	1.01 [0.85, 1.17]	2.84 [2.39, 3.30]	3.85 [3.45, 4.25]
$m = 2$, FIFO, $K = 20$	0.65 [0.51, 0.80]	2.63 [1.70, 3.56]	3.28 [2.39, 4.18]
$m = 2$, LIFO, $K = 0$	0.17 [0.09, 0.25]	4.30 [3.84, 4.76]	4.47 [4.01, 4.93]
$m = 2$, LIFO, $K = 20$	1.85 [1.59, 2.10]	3.78 [1.81, 5.75]	5.63 [3.81, 7.44]
$m = 3$, FIFO, $K = 0$	0.97 [0.81, 1.14]	2.20 [1.52, 2.87]	3.17 [2.48, 3.86]
$m = 3$, FIFO, $K = 20$	3.70 [3.54, 3.87]	1.84 [0.84, 2.85]	5.55 [4.54, 6.55]
$m = 3$, LIFO, $K = 0$	0.41 [0.31, 0.51]	3.86 [3.59, 4.13]	4.27 [4.01, 4.53]
$m = 3$, LIFO, $K = 20$	2.95 [2.77, 3.14]	3.06 [2.42, 3.70]	6.01 [5.40, 6.63]

Table 24: Full-horizon checked TabPFN versus checked myopic proposals. Profit differences are per period, with exploratory paired Student-t 95% intervals across ten independent seeds per cell. The last column counts post-warm-up action disagreements across ten pairs. Intervals are not simultaneous.

m	T	K	Issue	Mean difference	Paired 95% interval	Disagreements
2	31	0	FIFO	0.0000	[0.0000, 0.0000]	0
2	31	0	LIFO	0.0000	[0.0000, 0.0000]	0
2	31	20	FIFO	0.0000	[0.0000, 0.0000]	0
2	31	20	LIFO	0.0000	[0.0000, 0.0000]	0
3	15	0	FIFO	0.0000	[0.0000, 0.0000]	0
3	15	0	LIFO	0.0000	[0.0000, 0.0000]	0
3	15	20	FIFO	0.0000	[0.0000, 0.0000]	0
3	15	20	LIFO	0.0133	[−0.0168, 0.0435]	1

D.8 Operational sources of the primary profit gaps

The profit gap has different operational sources. In LS/FIFO, fewer shortages and higher sales offset the additional inventory cost. In BL/LIFO, sales revenue barely changes; lower ordering, holding, and expiration costs explain the improvement (Figure 3). These accounting decompositions identify service and inventory-cost trade-offs behind the comparison (Table 13); they do not isolate a single causal component of the pipelines.

E Planning Execution and Prior Sensitivity

E.1 Complete-horizon execution at small scale

Eight lost-sales cells use $m = 2, T = 31$ or $m = 3, T = 15$, crossed with FIFO/LIFO and $K = 0/20$, orders 0 or 1, and ten paired seeds per cell. Both checked policies use four common initial orders, zero terminal value, $H_{\max} = 2$, and the equal sequence mixture of Dirichlet(1) and Dirichlet$(1/d)$ priors. Demand alternates between the two laws in Appendix C. Exact rational Bellman values extend to the actual end of the experiment, with tolerance $1/t$. TabPFN inference is float32 with TF32 disabled; this is separate from the mixed-precision study below.

All 160 trajectories completed. Independent replay covered 11,040 transitions, including oracle trajectories, and all 3,040 post-warm-up checks passed. The two policies chose identical actions on 79 of 80 paired trajectories; one pair had one tolerated difference. The largest individual and cumulative gap was approximately 0.018613. Table 24 shows no demonstrated TabPFN profit benefit. Median trajectory times were 34.05/20.31 seconds for TabPFN/myopic proposals at $m = 2$, and 9.58/3.16 at $m = 3$, on a shared RTX 4090 host. The solver constructs exact values before checking the proposal, so this test verifies complete-horizon execution, not search acceleration.

A development case with $m = 3, T = 31$ exceeded 60 seconds and remains recorded as a failure. The separate original float32 two-period timing and cold-process memory probes are retained in the planning evidence archive. They showed no TabPFN acceleration and are not pooled with the mixed-precision measurements below.

E.2 Block-planning protocol

The frozen study contains 1,160 tasks and ten new paired seeds, 260926200–260926209. The main matrix has four LS cells (FIFO/LIFO, $K = 0/20$), $m = 3$, orders 0–600, $T = 10{,}000$, and the costs and ten demand families in Appendix B.4. Each cell has nine configurations: AC, structured FO, checked TabPFN blocks, and both myopic and myopic-ordered blocks under three priors. The priors are Dirichlet(1), Dirichlet$(1/d)$, and their equal *sequence* mixture; checked TabPFN uses the last. Each learner knows the physical model and period upper bound $H_{\max} = 10$, but not the demand laws or phase labels.

The small matrix crosses $m = 2/3$, FIFO/LIFO, and $K = 0/20$, with orders 0–8, $T = 1{,}000$, and the two demand laws in Appendix C. The same nine learners and an oracle task run per cell and seed. Each oracle task yields both dynamic and myopic trajectories with known laws and phase, empty initial stock, and no warm-up. Thus there are 360 main tasks, 800 small tasks, and 1,240 trajectories. Learners share demand paths and four initial orders within cell and seed; profits include initialization.

Checked policies optimize the remaining one or two periods in predetermined blocks, with zero terminal value and posterior updates on every hypothetical branch. Their tolerance is $1/t$ at one-based period t, starting at period five. Bounds cover all actions, including those not yet refined. No random action is drawn after certification. These checks bound model-internal block gaps; Eq. (A.24) still charges block boundaries.

Precision and inference. TabPFN 2.0.9 uses one estimator, internal temperature 0.5, random state zero, native float16 autocast, and disabled TF32. The planner retains float64 and directed high-precision arithmetic. This replaced an unfinished float32 study for runtime reasons before its test profits were inspected. Its 112 completed and 98 interrupted tasks remain archived and excluded. The separate complete-horizon float32 study above is retained independently.

All tasks finished before test-profit analysis. The primary family consists of four contrasts per main cell: checked TabPFN minus AC, structured FO, same-prior myopic control, and same-prior myopic-ordered blocks. The 16 two-sided paired tests use Holm correction; Student-t 95% intervals are pointwise. The 76 prior and small-instance contrasts are exploratory. The numerical reporting tolerance of 0.00000001 per seed does not replace raw differences or an equivalence test.

E.3 Block-planning results

Audit and primary comparisons. All tasks passed independent replay of 4,320,000 learner transitions, 160,000 oracle transitions, and 3,356,640 certificates. Demands and initialization match within all 120 paired cell–seed groups. Independent Bellman calculations verify the small-instance oracle values and actions. These checks validate the executions, not true-environment optimality.

Table 25 retains all 16 primary comparisons. Checked TabPFN blocks lose to AC, structured FO, and same-prior myopic control in every main cell; the largest Holm-adjusted p among these twelve contrasts is 0.00001678. Proposal-ordering differences are below 0.001 with all intervals containing zero, giving no demonstrated gain or equivalence. One lower endpoint rounds to 0.0000 but is negative before rounding (approximately -0.00004253).

Prior mass changes control performance. Table 26 retains every main configuration, including unsuccessful priors. Uniform Dirichlet(1) assigns 1,801 pseudo-observations per candidate residue, whereas Dirichlet$(1/d)$ assigns total mass one. For the sparse prior, blocks improve on its myopic policy by 6.66 $[4.97, 8.36]$ and 9.39 $[7.54, 11.25]$ in FIFO, but lose 6.30 $[5.11, 7.49]$ and 6.36 $[4.99, 7.73]$ in LIFO. These exploratory results do not establish TabPFN benefits: no sparse-prior TabPFN proposer was evaluated.

Sequence-likelihood dominance does not imply componentwise finite-profit dominance. The mixture retains its theoretical likelihood bound, while the large alphabet and short blocks leave conservative prediction and boundary charges. Our experiment does not separately estimate the contribution of prior error and terminal-value error to realized profit.

A structured demand baseline from the retained records. We additionally compare sparse-prior myopic control with AC and structured FO on the same ten seeds in each main cell. This

Table 25: Checked TabPFN minus each primary control in the four 10,000-period main cells. Differences are all-period mean profit, with paired Student-t 95% intervals across ten seeds. Two-sided p values use Holm correction across all 16 comparisons. A dagger marks numerical equality at absolute tolerance 0.00000001 for every paired seed; raw values remain in the source data. This is not an equivalence test.

Cell	Control	Difference	Paired 95% CI	Holm p
FIFO, $K = 0$	AC	-354.8904	[-381.3353, -328.4455]	< 0.0001
FIFO, $K = 0$	Structured FO	-347.2179	[-374.0796, -320.3562]	< 0.0001
FIFO, $K = 0$	Mixture myopic	-17.0599	[-20.1941, -13.9256]	< 0.0001
FIFO, $K = 0$	Myopic-ordered block	0.0001	[-0.0003, 0.0006]	1.0000
LIFO, $K = 0$	AC	-410.1667	[-438.8509, -381.4826]	< 0.0001
LIFO, $K = 0$	Structured FO	-406.2083	[-433.1895, -379.2272]	< 0.0001
LIFO, $K = 0$	Mixture myopic	-17.7079	[-18.9327, -16.4830]	< 0.0001
LIFO, $K = 0$	Myopic-ordered block	-0.0002	[-0.0008, 0.0004]	1.0000
FIFO, $K = 20$	AC	-347.8176	[-376.4702, -319.1650]	< 0.0001
FIFO, $K = 20$	Structured FO	-343.2902	[-370.1562, -316.4243]	< 0.0001
FIFO, $K = 20$	Mixture myopic	-14.7241	[-18.0293, -11.4189]	< 0.0001
FIFO, $K = 20$	Myopic-ordered block	0.0003	[-0.0006, 0.0011]	1.0000
LIFO, $K = 20$	AC	-406.5342	[-442.1703, -370.8981]	< 0.0001
LIFO, $K = 20$	Structured FO	-404.7773	[-431.7419, -377.8126]	< 0.0001
LIFO, $K = 20$	Mixture myopic	-17.7370	[-19.0507, -16.4232]	< 0.0001
LIFO, $K = 20$	Myopic-ordered block	0.0003	[0.0000, 0.0005]	0.3409

Table 26: All-period mean profit across ten independent seeds per cell, main test, $m = 3$. Values are descriptive cell means; paired uncertainty is reported in the comparison tables.

Policy	Demand prior	FIFO, $K = 0$	LIFO, $K = 0$	FIFO, $K = 20$	LIFO, $K = 20$
AC	–	378.76	351.26	358.50	331.56
Structured FO	–	371.09	347.30	353.97	329.80
Checked TabPFN	Sequence mixture	23.87	-58.91	10.68	-74.98
Mixture myopic	Dirichlet(1)	-2415.33	-2802.75	-2435.11	-2815.66
Myopic-ordered block	Dirichlet(1)	-2962.26	-3208.49	-2975.34	-3216.25
Mixture myopic	Dirichlet$(1/d)$	418.02	365.13	402.99	349.67
Myopic-ordered block	Dirichlet$(1/d)$	424.68	358.83	412.39	343.31
Mixture myopic	Sequence mixture	40.93	-41.20	25.41	-57.24
Myopic-ordered block	Sequence mixture	23.87	-58.91	10.68	-74.98

post hoc eight-comparison family is separate from the original 16 primary and 76 exploratory contrasts. All eight paired mean differences are positive after Holm correction (largest adjusted $p = 0.01338$; Table 28). Exhaustive paired sign-flip sensitivity gives adjusted $p \leq 0.015625$ under the corresponding symmetry assumption. The period-mixture learner receives $H_{\max} = 10$ and the clock, so these are structured-pipeline comparisons with different supplied information. The script, input fingerprint, seed differences, and all tests are retained in `sparse_prior_reanalysis/`. The FO arm here is the fixed structured rule, not the independently selected FO pipeline from Table 11.

Small-instance behavior. For sequence-mixture blocks versus the same-mixture myopic policy, five exploratory intervals are positive, one negative, and two contain zero (Table 29). Checked TabPFN falls below the dynamic oracle in every cell. All other arm means, oracle contrasts, and the complete 76-comparison family remain in `arm_summary.csv` and `secondary_paired.csv`. The full tables are also retained as archived source.

On these small instances, mean conditional KL for uniform/sparse/sequence priors is 0.1117/0.0418/0.0422 at $m = 2$, and 0.1637/0.0468/0.0472 at $m = 3$. These known-law diagnostics are not benchmark KL estimates and do not isolate the benchmark profit loss. Conditional cross-entropy and realized prequential log loss remain separately reported in the data.

Table 27: Exploratory paired profit differences, main test, $m = 3$. Each comparison uses ten seeds; Student-t 95% intervals are pointwise and are not simultaneous. A dagger has the numerical-equality meaning of Table 25.

Cell	Comparison	Difference	Paired 95% CI
FIFO, $K = 0$	Block minus myopic, Dirichlet(1)	-546.9320	[-548.4387, -545.4252]
FIFO, $K = 0$	Block minus myopic, Dirichlet$(1/d)$	6.6610	[4.9663, 8.3557]
FIFO, $K = 0$	Block minus myopic, Sequence mixture	-17.0600	[-20.1940, -13.9260]
FIFO, $K = 0$	Dirichlet$(1/d)$ minus Dirichlet(1), block	3386.9410	[3379.6122, 3394.2698]
FIFO, $K = 0$	Sequence mixture minus Dirichlet(1), block	2986.1325	[2960.6467, 3011.6183]
LIFO, $K = 0$	Block minus myopic, Dirichlet(1)	-405.7417	[-407.1210, -404.3624]
LIFO, $K = 0$	Block minus myopic, Dirichlet$(1/d)$	-6.3001	[-7.4907, -5.1095]
LIFO, $K = 0$	Block minus myopic, Sequence mixture	-17.7076	[-18.9321, -16.4831]
LIFO, $K = 0$	Dirichlet$(1/d)$ minus Dirichlet(1), block	3567.3259	[3559.7844, 3574.8674]
LIFO, $K = 0$	Sequence mixture minus Dirichlet(1), block	3149.5889	[3122.4158, 3176.7620]
FIFO, $K = 20$	Block minus myopic, Dirichlet(1)	-540.2234	[-542.6570, -537.7898]
FIFO, $K = 20$	Block minus myopic, Dirichlet$(1/d)$	9.3932	[7.5369, 11.2495]
FIFO, $K = 20$	Block minus myopic, Sequence mixture	-14.7244	[-18.0291, -11.4197]
FIFO, $K = 20$	Dirichlet$(1/d)$ minus Dirichlet(1), block	3387.7230	[3380.9048, 3394.5413]
FIFO, $K = 20$	Sequence mixture minus Dirichlet(1), block	2986.0158	[2959.7020, 3012.3296]
LIFO, $K = 20$	Block minus myopic, Dirichlet(1)	-400.5953	[-401.8963, -399.2944]
LIFO, $K = 20$	Block minus myopic, Dirichlet$(1/d)$	-6.3613	[-7.7322, -4.9903]
LIFO, $K = 20$	Block minus myopic, Sequence mixture	-17.7372	[-19.0509, -16.4235]
LIFO, $K = 20$	Dirichlet$(1/d)$ minus Dirichlet(1), block	3559.5628	[3551.9999, 3567.1257]
LIFO, $K = 20$	Sequence mixture minus Dirichlet(1), block	3141.2767	[3114.2373, 3168.3162]

Table 28: Exploratory reanalysis of sparse-prior myopic control in the block-planning study. Per-period profit differences over all 10,000 periods use the ten paired seeds. Intervals are pointwise 95% Student-t intervals. The demand mixture receives the period upper bound $H_{\max} = 10$; AC and structured FO retain their original inputs. All eight two-sided paired tests remain positive after a separate Holm correction.

Cell	Sparse myopic − AC	Sparse myopic − structured FO
FIFO, $K = 0$	$39.26\ [23.89, 54.63]$	$46.93\ [42.92, 50.94]$
LIFO, $K = 0$	$13.87\ [3.65, 24.10]$	$17.83\ [15.49, 20.16]$
FIFO, $K = 20$	$44.50\ [23.30, 65.69]$	$49.02\ [44.83, 53.21]$
LIFO, $K = 20$	$18.11\ [7.55, 28.67]$	$19.87\ [17.26, 22.48]$

Proposal cost. An isolated probe evaluates 24 fixed validation states with five proposal orderings and four budgets (480 calls). Median fixed-quality total times, including setup, are 0.5605/0.5567/0.5352/0.5894/0.7248 seconds for fixed/random/myopic/XGBoost/TabPFN, with all decisions certified. No ordering meets deadline and certificate at 0.05 or 0.25 seconds; all do at 1.00 second. Correlated-state timings are descriptive; shared-host measurements remain separate.

Table 29: Small-instance sequence-mixture contrasts. Mean per-period differences and exploratory paired 95% intervals across ten seeds. *Block* uses myopic proposal ordering; *checked* uses TabPFN proposals. Learner–oracle gaps also contain information and initialization differences.

m	Cell	Block − myopic	Checked − dynamic oracle
2	FIFO, $K = 0$	$0.50\ [0.26, 0.74]$	$-0.34\ [-0.40, -0.27]$
2	LIFO, $K = 0$	$-0.14\ [-0.31, 0.02]$	$-0.63\ [-0.75, -0.51]$
2	FIFO, $K = 20$	$-0.71\ [-0.95, -0.47]$	$-1.37\ [-1.63, -1.11]$
2	LIFO, $K = 20$	$0.82\ [0.38, 1.25]$	$-1.63\ [-1.86, -1.40]$
3	FIFO, $K = 0$	$0.42\ [0.18, 0.65]$	$-0.40\ [-0.47, -0.32]$
3	LIFO, $K = 0$	$0.08\ [-0.06, 0.21]$	$-0.57\ [-0.70, -0.45]$
3	FIFO, $K = 20$	$1.34\ [1.17, 1.51]$	$-2.29\ [-2.52, -2.05]$
3	LIFO, $K = 20$	$2.43\ [2.27, 2.59]$	$-0.54\ [-0.94, -0.14]$

F Original Eight-Regime Benchmark

This earlier benchmark uses ten runs, lifetime three, and all 10,000 periods. Its uncertainty is $1.96s/\sqrt{10}$ across run means. Classical and cyclic references are selected in hindsight; they are not dynamic-optimum upper bounds. The benchmark uses an earlier comparator set, whereas the independently selected same-backbone FO test in the main text is the relevant interface comparison. Decision-XGB here keeps the native action-sampling rule with 50 trees, depth five, and learning rate

Table 30: Full-horizon mean profits over ten runs, with 95% intervals, including Decision-XGB under the same action-scoring interface. Classical and cyclic references are tuned in hindsight; cyclic references also use the true period. Marginal intervals do not establish paired significance. Every policy is evaluated over all 10,000 periods.

Policy	LS/FIFO	LS/LIFO	BL/FIFO	BL/LIFO
Setup cost $K = 0$				
Decision-TabPFN	395.29 ± 9.76	342.87 ± 13.40	554.26 ± 4.02	529.35 ± 6.21
Decision-XGB	367.65 ± 10.69	273.13 ± 35.02	516.93 ± 6.87	461.82 ± 13.27
Meta-HQL	276.30 ± 44.13	218.13 ± 53.13	391.98 ± 102.24	373.92 ± 101.96
TS	-43.08 ± 36.52	227.71 ± 21.01	542.26 ± 1.62	349.00 ± 3.99
XGBoost-FO	269.99 ± 21.78	295.97 ± 6.25	526.50 ± 2.22	323.28 ± 5.78
Base stock	331.08 ± 2.68	149.60 ± 2.50	531.22 ± 1.48	501.32 ± 1.42
Cyclic base stock	474.36 ± 3.74	379.85 ± 4.04	581.25 ± 1.50	567.99 ± 1.51
Setup cost $K = 20$				
Decision-TabPFN	373.80 ± 12.85	316.82 ± 6.80	539.79 ± 3.09	515.74 ± 2.12
Decision-XGB	338.05 ± 17.33	256.60 ± 21.66	488.86 ± 16.02	445.51 ± 9.45
Meta-HQL	260.60 ± 44.65	202.03 ± 53.59	375.08 ± 102.78	356.77 ± 102.46
TS	-62.63 ± 36.57	208.44 ± 21.04	524.02 ± 1.61	330.02 ± 4.00
XGBoost-FO	250.96 ± 21.88	276.90 ± 6.27	508.81 ± 2.20	304.79 ± 5.76
(s, S)	311.09 ± 2.66	132.04 ± 5.24	511.53 ± 1.47	490.42 ± 1.83
Cyclic (s, S)	458.38 ± 3.66	365.46 ± 3.96	564.57 ± 1.52	547.81 ± 1.54

0.1. The controlled common-query XGBoost model instead uses the configuration in Appendix C. The source archive retains the longer-lifetime plots, backbone and component ablations, and earlier exploratory FO sweeps with their separate protocols. These records do not replace the stronger independently selected FO comparison.

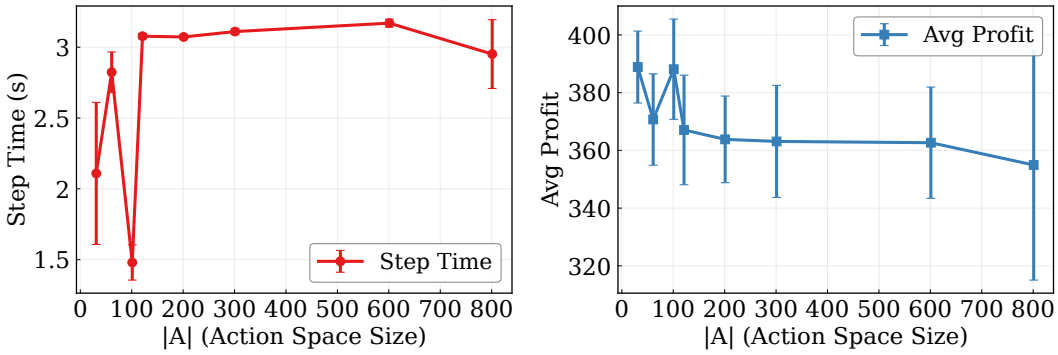

Figure 4: **Original action-grid probe.** Per-step inference time and mean daily profit versus the number of candidate actions; error bars are the source study's standard deviations across runs. This measures the original scorer under that protocol, not the later checked planner or seven-GPU throughput.

F.1 SPOILAGE TRAJECTORIES

Figure 5 complements the eight profit trajectories in Figure 1. Curves begin at period 50 for visibility; each cumulative statistic includes the preceding periods. Shading is the mean plus or minus 1.96 standard errors over ten runs. XGB denotes XGBoost-FO. Lower spoilage can reflect under-ordering and must be read together with profit.

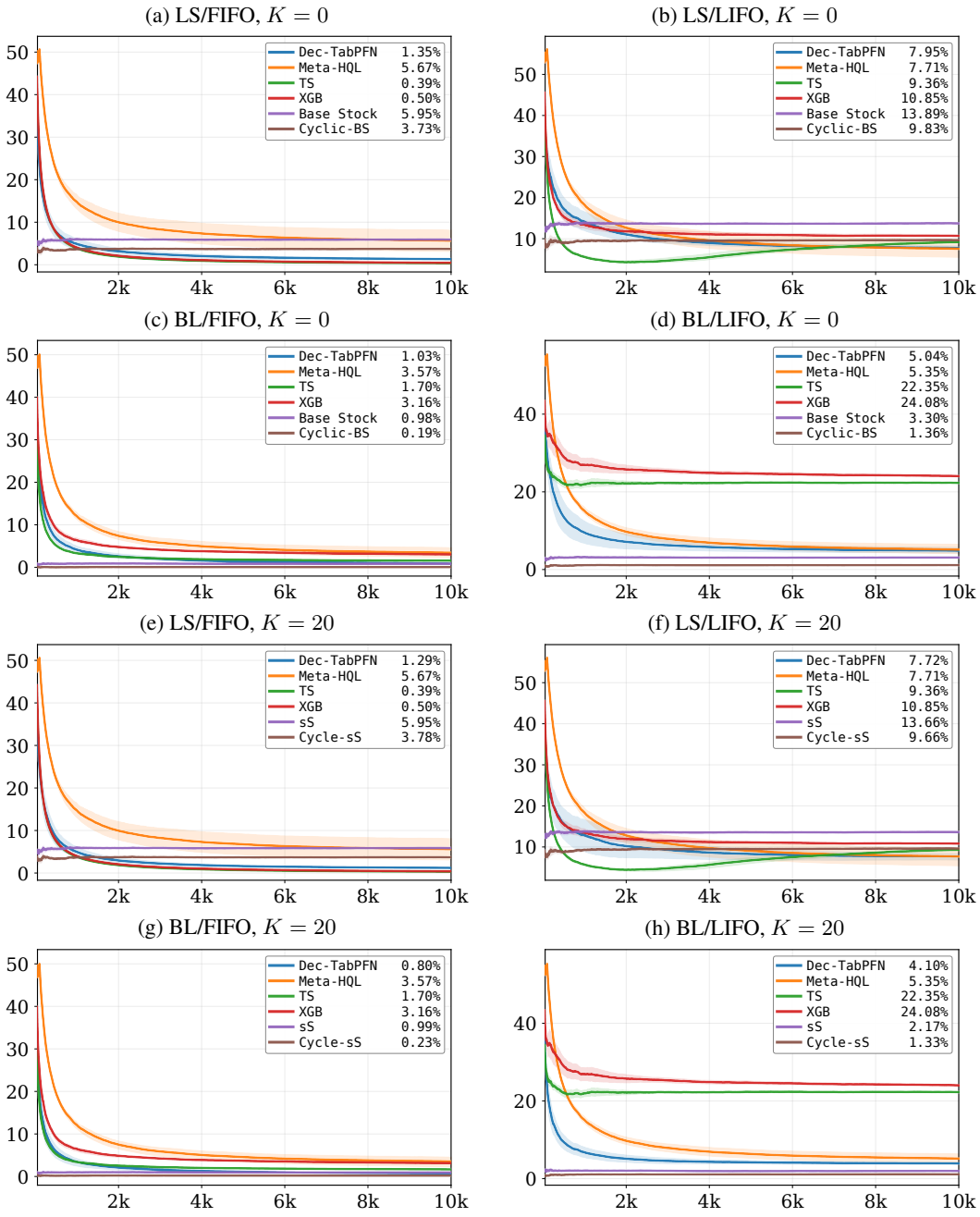

Figure 5: **Profit and waste need not have the same ranking.** Cumulative discarded units as a percentage of cumulative orders, versus period. Ten-run means and pointwise 95% intervals; $m = 3$, $H = 10$, $T = 10{,}000$. Legends report the final cumulative fractions.

F.2 Cycle-position diagnostics

Figures 6–7 show how orders align with demand over the ten positions of the hidden cycle. Early, middle, and late windows divide the 10,000-period trajectory into 3,334/3,333/3,333 periods. Correlations are computed across cycle-position means. They are descriptive aggregate diagnostics: neither alignment nor the within-profile correlation p-values establish an internal phase representation or provide independent-seed policy comparisons.

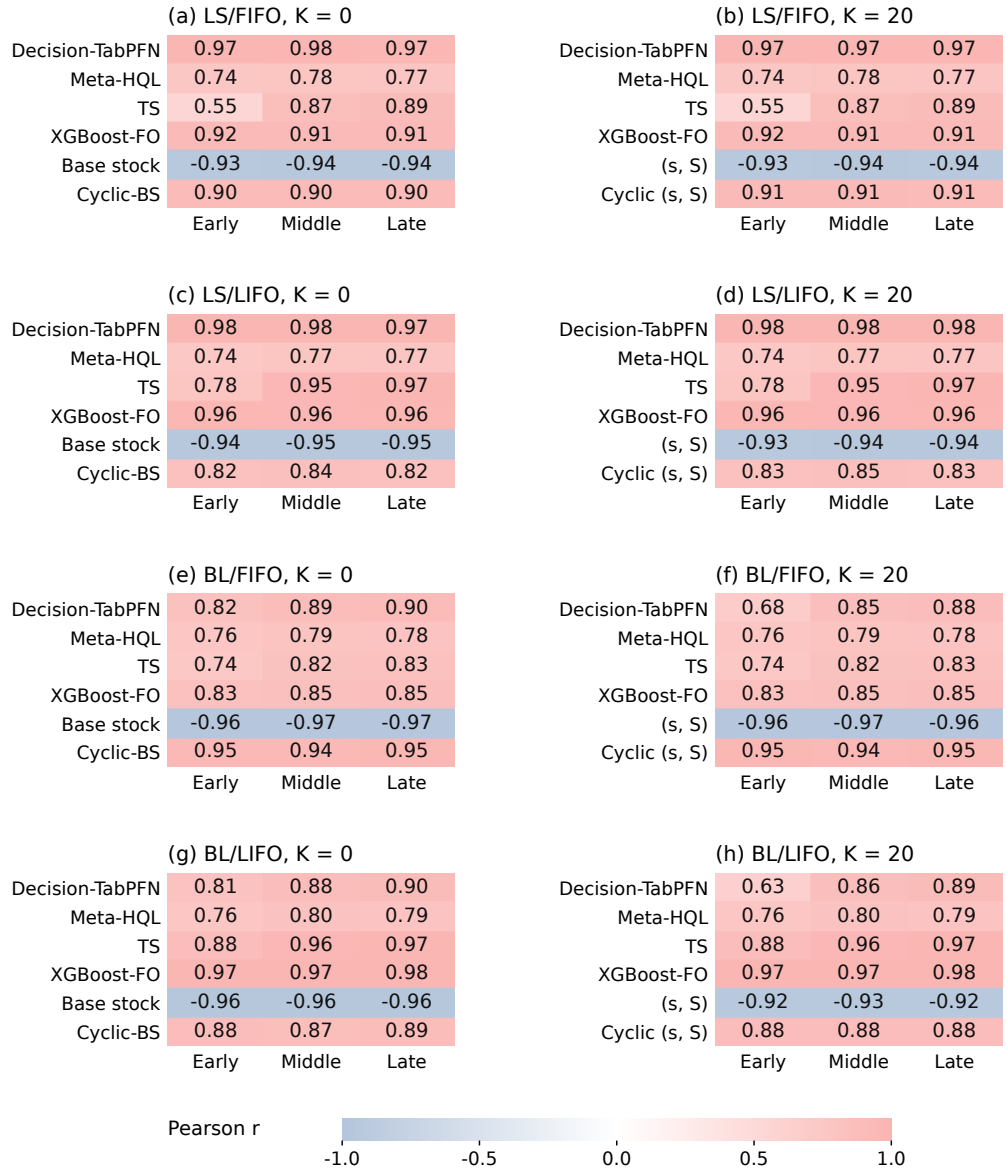

Figure 6: **Cycle alignment across all eight regimes.** Columns distinguish $K = 0$ and $K = 20$; rows distinguish LS/FIFO, LS/LIFO, BL/FIFO, and BL/LIFO. Each panel retains all six policies and the early, middle, and late horizon thirds from the original heatmap. Cells show Pearson correlations between cycle-position mean orders and demand, with one shared color scale. Correlation measures timing, not profitability or dynamic optimality.

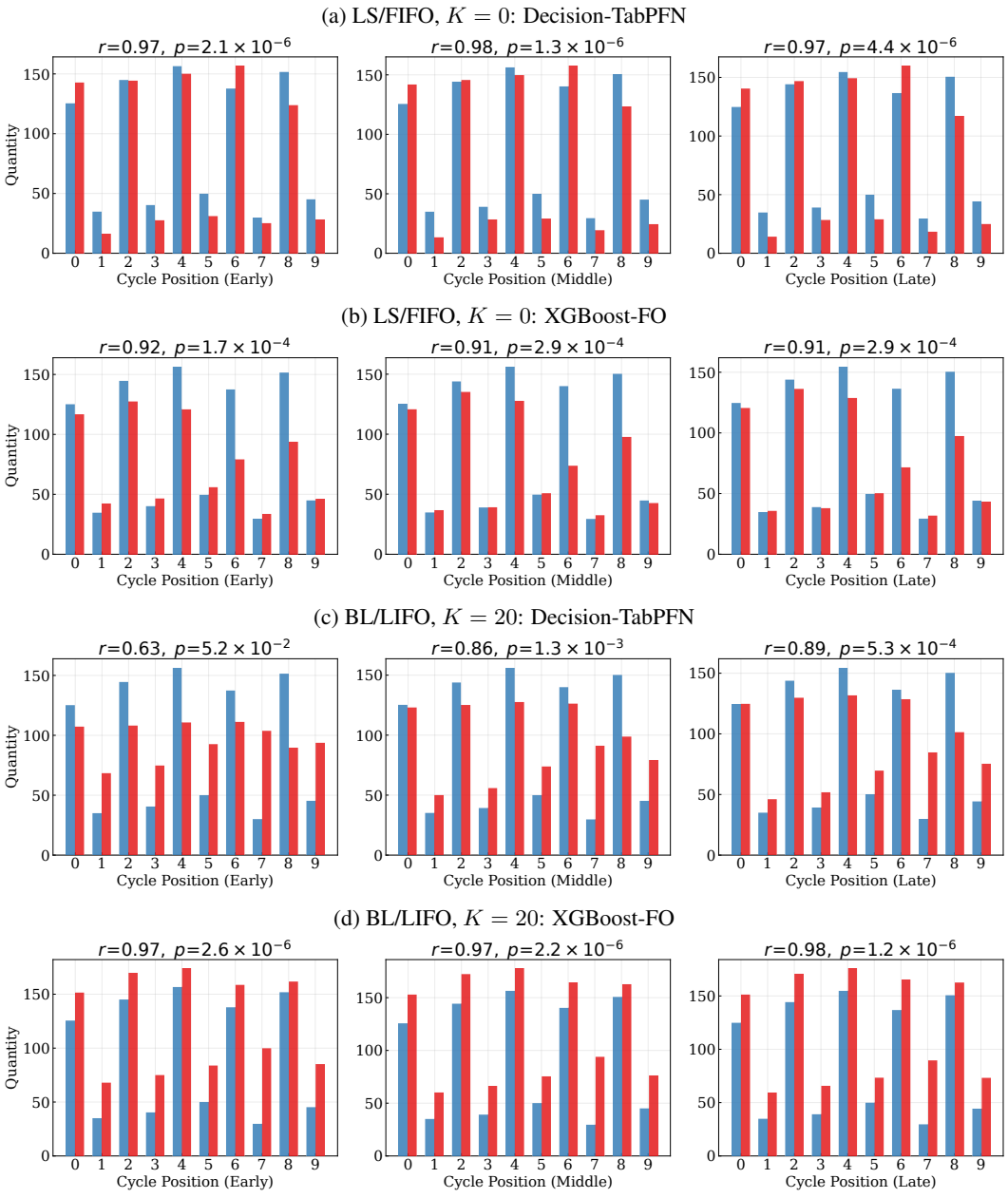

Figure 7: **Order levels complement the correlation summaries.** Two representative regimes from Figure 1, with Decision-TabPFN and the historical XGBoost-FO baseline. Within each row, panels show early, middle, and late cycle profiles; blue bars are mean orders and red bars are mean demand. Profile correlations are descriptive across the ten cycle positions.

F.3 ORIGINAL NON-PERISHABLE TRANSFER PANELS

The synthetic and sales-driven comparisons below use Meta-HQL's native non-perishable environment. They retain the original protocol and reference labeled OPT. They are separate from the perishable controlled suite and do not support a claim of optimality for Decision-TabPFN.

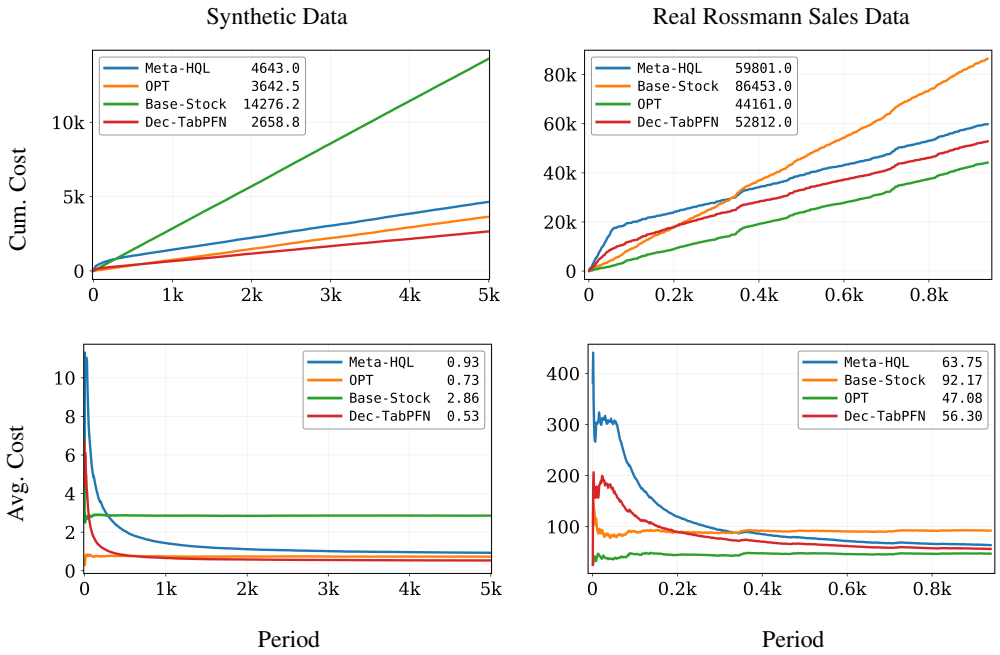

Figure 8: Inventory control under Meta-HQL's non-perishable environment. Rows: cumulative (top) and average (bottom) cost; columns: synthetic (H=5, left) and Rossmann (H=7, right). Methods: Decision-TabPFN (ours), Meta-HQL (Gong & Simchi-Levi, 2024), OPT (reference labeled in the source experiment), and Base-Stock.