Autoregressive Decision Making from Observable Histories

Yunbei Xu[1] Yuzhe Yuan[1] Ruohan Zhan[2]
[1]National University of Singapore [2]University College London
yunbei@nus.edu.sg e1374511@u.nus.edu ruohan.zhan@ucl.ac.uk

Abstract

Sequential operational decisions are made from observable histories, yet common pipelines compress them into a presumed Markov state, imitate logged actions, or predict objective-specific values. We instead learn the *controlled observable continuation law* $q^\star(z_{t+1} \mid h_t, a_t)$ and optimize under the learned law.

Four distinct sources drive the advantage. First, full observable history avoids irreversible compression: an exact decomposition shows that the best compressed KL risk equals a conditional mutual information. Second, a correct operational reactor maps shocks into observables, enforcing feasibility, contracting KL, and preserving log-ratio cover complexity. Third, joint-likelihood learning reveals exactly where horizon enters: a constructed decomposable class forces linear-in-H estimation even for unrestricted trajectory-law estimators, whereas the analogous fully shared lower bound holds only for proper selectors because improper aggregation is horizon-free. ERM upper bounds match the relevant H/n scale up to logarithmic and boundedness factors, while the exact KL chain rule adds no horizon multiplier to best-in-class joint-KL approximation. Fourth, the controlled law is reusable across objectives: bounded objectives have a tight square-root-KL guarantee, while strongly curved, gradient-stable programs admit regret linear in KL. Of particular importance is the role of $H-$round trajectory KL in reducing sensitivity to realizability.

Mechanism-focused experiments recover these predictions. Lost-sales inventory and reflected-queue experiments demonstrate the representation gains from observable histories and the feasibility benefits of structural reactors. Pricing experiments identify the square-root and linear KL-to-regret regimes under near ties and strong curvature, respectively. Finally, a 40-store semi-synthetic Rossmann study reproduces the predictive and feasibility patterns under realistic serial and calendar variation, while achieving a 33.6% reduction in constructed one-step cost relative to the logging policy on held-out states through safe, valid autoregressive learning.

Keywords: autoregressive learning; observable histories; sequential decision making; trajectory KL; generative modeling

Contents

1 Introduction

Operational systems are observed through histories. A pricing platform records prices, sales, promotions, stockouts, and calendar effects. An inventory system records orders, censored sales, replenishments, and supplier events. A service system records arrivals, departures, abandonment, workload, and staffing. These histories are high-dimensional and partially observed; a useful state may exist, but it is often unknown and need not be sufficient for the next operational consequence.

The resulting learning problem is not simply to predict the next element of a log. Actions are interventions, and logged actions may encode legacy rules, human judgment, changing objectives, or unobserved constraints. Imitating them learns the behavior rule that generated the data. Operational decision making instead asks what will happen under each candidate intervention. We therefore take as primitive

$$q_t^\star(z_{t+1} \mid h_t, a_t), \tag{1}$$

the law of the next observable consequence z_{t+1} given the observable history h_t and candidate action a_t. The action is an input, not the label. The decision pipeline is

$$\text{learn the continuation law} \quad \longrightarrow \quad \text{optimize under that law} \quad \longrightarrow \quad \text{act.}$$

This is a *generative/autoregressive* extension of the established predict-then-optimize (or "end-to-end" learning) paradigm in the discriminative setting (Donti et al., 2017; Elmachtoub and Grigas, 2022; Lagzi et al., 2026). Instead of learning a single conditional mean or task-specific value function, it learns the controlled observable continuation distribution, yielding a reusable generative model that supports multiple downstream objectives and constraints.

Why observable histories? The full observable history is a principled default representation when a sufficient state is not known. We prove an exact identity: compressing h_t to $s_t = \phi(h_t)$ incurs an irreducible one-step KL gap equal to a conditional mutual information, and the gap vanishes exactly when s_t is conditionally sufficient on the covered history–action distribution. Observable histories also admit domain structure. A known reactor can map latent demand or net input into observable sales, inventory, or workload; the resulting model obeys feasibility by construction, contracts KL by data processing, and does not enlarge log-ratio covers. Thus history and structure are complements: history avoids an unjustified state compression, while the reactor avoids asking a neural decoder to relearn operational identities.

The technical backbone. Conditional log loss is the joint trajectory log likelihood. For a decomposable class $\mathcal{Q}_0^H$ with $|\mathcal{Q}_0| = M$, let G bound the relevant conditional log-likelihood ratios. Our strengthened lower bound constructs a realizable family for which every learner, including an unrestricted improper trajectory-law estimator, pays

$$\mathbb{E}\mathrm{KL}(P^\star \| \widehat{P}) \gtrsim \frac{HG^2 \log M}{n}.$$

A separate construction uses one conditional kernel at every round and gives the same $HG^2 \log M/n$ lower-bound order for proper selectors over a fully shared class. This properness qualifier cannot be removed: for any M candidate trajectory laws, posterior aggregation attains expected excess KL at most $\log M/(n+1)$, independently of H. The resulting lower-bound messages and the

corresponding restrictions are therefore substantially different. Conversely, log-loss ERM satisfies, with high probability,

$$\mathrm{KL}(P^\star \| P_{\hat{q}}) \le (1+\varepsilon) \inf_{q \in \mathcal{Q}_0^H} \mathrm{KL}(P^\star \| P_q) + \widetilde{O}\left(\frac{H(G+1)\log M}{\varepsilon n} \right).$$

Here $\widetilde{O}$ suppresses logarithmic factors in the displayed complexity and confidence parameters. The lower and upper bounds match in their linear horizon and inverse sample-size dependence, up to logarithmic and boundedness factors. The substantive point is the separation they reveal: statistical error accumulates over the horizon, but the exact autoregressive KL chain rule introduces no additional factor of H in front of the best-in-class *joint*-KL approximation error. These are worst-case existence results, not statements that every class necessarily pays a linear horizon cost. The decomposable result ranges over arbitrary output laws; the fully shared result ranges only over proper class-valued selectors.

From learning to decisions and realizability sensitivity. Let Δ_D be the worst joint KL induced by the learned law over a supported decision class. For an objective of range B, plug-in regret is at most $B\sqrt{2\Delta_D}$, and this square-root dependence is sharp. If the downstream contextual stochastic-optimization objective is strongly concave and its gradient is stable under KL perturbations, the same model error yields the fast rate $O(\Delta_D)$. This gain belongs to the continuation-learning plus optimization pipeline: a direct policy has no remaining optimization layer whose curvature can be used, and a greedy value-to-action map is generally discontinuous without a margin or stability assumption. Coverage remains essential; accuracy on logging histories does not automatically control a policy that visits unsupported histories.

Theorem 7 makes the approximation issue sharp. It constructs a misspecified continuation class whose natural H-round trajectory-KL error κ is independent of H, although the induced policy regret and optimal value- and action-value-function errors are $\Omega(H\sqrt{\kappa})$. In a companion strongly concave pricing problem, decision regret is instead $\Theta(H\kappa)$. Thus small approximation error for the observation-history law need not be small approximation error for an unnormalized H-round policy or value objective. This is a concrete metric- and geometry-level benchmark for the call to develop scalable RL methods that are not sensitive to realizability—the "elephant in the room" identified by the Simons Institute workshop (Simons Institute, 2020). It is not an impossibility result for all realizability-robust algorithms: such a guarantee must specify its approximation metric, horizon normalization, coverage, and decision geometry.

Evidence. The experiments are designed to test rigorous mechanisms underlying representation, generalization, and approximation, rather than to serve as a forecasting leaderboard. Structural GRUs have exactly feasible support and lower observable KL than matched direct GRUs in lost-sales inventory and a controlled Lindley queue, including after a prespecified feasibility projection. Inventory regret is also lower with paired statistical separation; the smaller queue-regret differences are not resolved. A finite-class experiment recovers H/n scaling, and a common persistent-demand misspecification yields pricing transfer slopes 0.499 and 0.999 under discrete and strongly concave decisions. The semi-synthetic Rossmann study supplies realistic serial and calendar variation, but its policy comparison is deliberately limited to one-step decisions on the held-out logging-state distribution.

Organization and contributions. Our preliminary short version (Xu et al., 2026) develops the joint-KL oracle and minimax learning results, together with a basic Pinsker rollout implication,

in an autoregressive next-token formulation. Its initial arXiv version did not consistently make clear that the formal lower bound applies only to proper selectors, nor did it establish the claimed extension to a realizable, fully shared model class beyond the decomposable product setting. Its proof also used ambiguous base-class versus product-class cardinality notation. Theorem 2 below strengthens the decomposable result to arbitrary trajectory-law outputs, Theorem 3 gives a separate proper shared-class construction, and Proposition 2 shows why the shared qualifier is necessary. Appendix A records the remaining limitations explicitly. The present paper makes controlled observable-history laws the primary learning object and adds the compression, structural-reactor, coverage-aware decision transfer, likelihood–value/policy separation, and experimental results.

The paper has three conceptual layers, followed by experimental evidence. Section 2 develops observable continuations as a representation and establishes the exact costs and benefits of history compression and structural reactors. Section 3 gives the main lower and upper bounds in joint KL. Section 4 converts those guarantees into decision performance and identifies when fast transfer is possible. Section 5 tests the three empirical mechanisms—representation, estimation, and transfer. Related work and all proofs are deferred to the appendices so that the main text follows this conceptual and technical progression directly.

2 Why Learn from Observable Histories?

2.1 Controlled laws from observable histories

Let $z_t \in \mathcal{Z}_t$ denote the information observed at time t, and let $a_t \in \mathcal{A}_t(h_t)$ be a feasible action after history

$$h_t = (z_1, a_1, z_2, a_2, \ldots, a_{t-1}, z_t) \in \mathcal{H}_t.$$

The controlled continuation law is the stochastic kernel

$$q_t^\star(\cdot \mid h_t, a_t) \in \Delta(\mathcal{Z}_{t+1}). \tag{2}$$

No finite-dimensional Markov state is assumed. Over H decisions, the law of an observable trajectory under a logging rule μ factorizes as

$$P_\mu^\star(z_{1:H+1}, a_{1:H}) = p_1^\star(z_1) \prod_{t=1}^{H} \mu_t(a_t \mid h_t) q_t^\star(z_{t+1} \mid h_t, a_t). \tag{3}$$

The behavior factor μ_t may be irregular even when the continuation factor is structured. Our target is the second factor.

Two conditions delimit the interventional interpretation. First, $q_t^\star(\cdot \mid h, a)$ must be identified at the history–action pairs of interest, for example through randomized logging or sequential ignorability given the modeled history. Second, candidate interventions must be covered by the logging distribution. Without identification, the learned law is observational; without coverage, a logging-law guarantee need not transfer to a new decision rule. Section 4 states the quantitative coverage conversion explicitly.

Given n independent logged trajectories, a conditional model is trained by

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{nH} \sum_{i=1}^{n} \sum_{t=1}^{H} -\log q_{t,\theta}(z_{t+1}^{(i)} \mid h_t^{(i)}, a_t^{(i)}). \tag{4}$$

The observation z_{t+1} is the label. The action is a candidate intervention included in the conditioning argument.

2.2 History, action, and value are different learning objects

Action-token modeling learns $p(a_t \mid h_t)$ and is appropriate when the goal is imitation. Objective-value prediction learns a scalar $\mu(h, a)$ tied to a particular downstream criterion. Continuation modeling learns the distribution $q^\star(z_{t+1} \mid h_t, a_t)$ and retains the information needed to change the objective, impose a new risk constraint, or perform a different rollout. These objects can be combined in an algorithm, but they are not statistically or decision-theoretically interchangeable.

Figure 1 locates continuation learning between action imitation and objective-specific value prediction.

2.3 The representation and approximation value of history

The case for the full observable history should not be overstated: if a known compressed state is sufficient, no information is lost. The following identity states exactly what is lost otherwise.

Proposition 1 (Exact cost of compressing observable history)**.** *Let (H, A, Z) take values in standard Borel spaces, let $S = \phi(H)$ be any measurable history representation, and let regular conditional distributions exist. For every kernel $r(dz \mid s, a)$ for which the displayed quantities are finite,*

$$\begin{aligned}&\mathbb{E}\big[\mathrm{KL}\big(P_{Z|H,A}(\cdot \mid H, A) \,\|\, r(\cdot \mid S, A)\big)\big] \\ &\quad = I(Z; H \mid S, A) + \mathbb{E}\big[\mathrm{KL}\big(P_{Z|S,A}(\cdot \mid S, A) \,\|\, r(\cdot \mid S, A)\big)\big]\,. \end{aligned} \tag{5}$$

Consequently,

$$\inf_r \mathbb{E}\mathrm{KL}(P_{Z|H,A} \| r_{Z|S,A}) = I(Z; H \mid S, A),$$

with minimizer $r = P_{Z|S,A}$. The compression is lossless precisely when $Z \perp H \mid (S, A)$, that is, when S is conditionally sufficient for the next observable consequence under the action.

This result is an approximation statement, not a claim that longer raw inputs always improve finite-sample performance. It is relative to the joint law of (H, A, Z); narrow action support can mask counterfactual insufficiency. Full history removes the irreducible KL gap created by an insufficient state on the covered distribution, while architecture, regularization, and sample size determine whether a learner can exploit that information.

2.4 Operational structure as a representation

Many systems also contain a known observable reactor

$$Z_{t+1} = \Phi_t(h_t, a_t, \xi_{t+1}), \qquad \xi_{t+1} \sim p_t^\star(\cdot \mid h_t, a_t), \tag{6}$$

so that $q_t^\star = \Phi_t(h_t, a_t, \cdot)_\# p_t^\star$. Lost-sales balance, Lindley reflection, capacity constraints, and sales censoring are canonical examples. A structural learner estimates the shock law p_t and retains Φ_t as a model layer. For mutually absolutely continuous kernels, define the uniform log-ratio distance by

$$d_{\log}(p, p') := \sup_x \left\| \log \frac{dp(\cdot \mid x)}{dp'(\cdot \mid x)} \right\|_{L_\infty(p'(\cdot|x))}, \tag{7}$$

and set it to $+\infty$ otherwise.

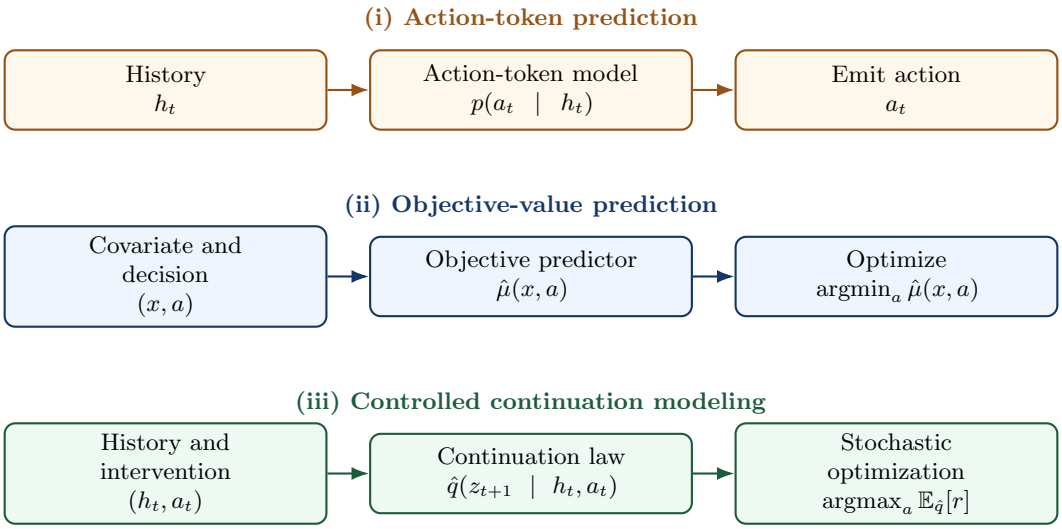

Figure 1: Three different primitives. Action-token prediction emits a decision. Objective-value prediction learns a scalar criterion and optimizes it. Controlled continuation modeling learns the law of observable consequences under candidate decisions and then solves an explicit stochastic optimization problem.

Theorem 1 (Structural feasibility and KL contraction)**.** *Fix a horizon H. Suppose the true continuation law admits the structural representation*

$$q_t^\star(\cdot \mid H_t, A_t) = \Phi_t(H_t, A_t, \cdot)_{\#} p_t^\star(\cdot \mid H_t, A_t), \qquad t = 1, \dots, H.$$

Assume each map $\Phi_t(h, a, \cdot)$ is measurable and its image $\mathcal{S}_t(h, a)$ below is measurable. All conditional assertions are understood at the relevant histories, or $P_d^\star$-almost surely for the trajectory statement. Then the following hold.

1. ***Constraint satisfaction.*** *Let*

   $$\mathcal{S}_t(h, a) := \operatorname{Im}\{\Phi_t(h, a, \cdot)\}$$

   be the feasible observable set generated by the reactor. Every $q_t^p \in \mathcal{Q}_{\Phi,t}$ satisfies

   $$q_t^p(\mathcal{S}_t(h, a) \mid h, a) = 1.$$

   Hence generated continuations automatically obey the balance, reflection, censoring, or capacity constraints encoded by Φ_t.

2. ***KL contraction.*** *For any candidate shock law $p_t \in \mathcal{P}_t$,*

   $$\mathrm{KL}(q_t^\star(\cdot \mid h, a) \,\|\, q_t^p(\cdot \mid h, a)) \le \mathrm{KL}(p_t^\star(\cdot \mid h, a) \,\|\, p_t(\cdot \mid h, a)).$$

   Consequently, for any fixed possibly history-dependent decision object d, provided the true and candidate trajectory laws use the same initial law and the same decision kernel,

   $$\mathrm{KL}(P_d^\star \| P_d^{q^p}) \le \sum_{t=1}^{H} \mathbb{E}_{P_d^\star} \mathrm{KL}(p_t^\star(\cdot \mid H_t, A_t) \,\|\, p_t(\cdot \mid H_t, A_t)).$$

3. ***Log-ratio cover nonexpansion.*** *If $\mathcal{P}_t$ has an r-cover in uniform log-ratio distance (7) with size $N_{\log}(\mathcal{P}_t, r)$, then the induced continuation class satisfies*

$$N_{\log}(\mathcal{Q}_{\Phi,t}, r) \le N_{\log}(\mathcal{P}_t, r).$$

Thus this log-ratio covering complexity of the structural continuation class is no larger than that of the shock-law class.

The theorem separates three benefits. Feasibility is representational: the model class cannot emit a continuation outside the reactor image. KL contraction is an approximation guarantee: an error in the shock law cannot become larger after observation. Cover nonexpansion is statistical: encoding the reactor does not add log-ratio covering complexity. None of these statements says that every structural learner dominates every direct learner; they identify what is gained when the operational map is correct.

Table 1 records how the same learning object specializes across operational domains.

3 Joint KL: Statistics and Approximation

3.1 Joint likelihood and the exact chain rule

We now state the statistical problem without operational notation. Let $U^1_{1:H}, \dots, U^n_{1:H} \overset{\text{iid}}{\sim} P^\star$, where

$$P^\star(u_{1:H}) = \prod_{h=1}^{H} q^\star_h(u_h \mid u_{<h}), \qquad P_q(u_{1:H}) = \prod_{h=1}^{H} q_h(u_h \mid u_{<h}).$$

For

$$K_h(q_h) := \mathbb{E}_{U_{<h} \sim P^\star_{<h}} \mathrm{KL}(q^\star_h(\cdot \mid U_{<h}) \| q_h(\cdot \mid U_{<h}))\,, \tag{8}$$

the autoregressive chain rule gives the identity

$$\mathrm{KL}(P^\star \| P_q) = \sum_{h=1}^{H} K_h(q_h). \tag{9}$$

Moreover, population log loss equals a constant plus this joint KL. Training, sequence-level evaluation, and best-in-class approximation are therefore measured in the same divergence.

We distinguish the decomposable class $\mathcal{Q}_0^H$, whose conditional component may vary with h, from the fully shared class $\mathcal{Q}_{\text{sh}} = \{(q_0, \dots, q_0) : q_0 \in \mathcal{Q}_0\}$. The former isolates the intrinsic statistical cost of estimating H conditional coordinates, whereas the latter captures complete parameter sharing while allowing the horizon dependence to reflect problem-specific effective replication. Here, "fully shared" means that the same indexed history kernel is used at every round, although the kernel may condition on both the initial observable context and the evolving history.

These structural labels are distinct from learner scope. Following the proper/improper terminology used by Rohatgi et al. (2025), a *proper* estimator must return one member of the candidate class, whereas an *improper* estimator may return any trajectory law, including a mixture of complete candidate laws. Their "no parameter sharing" condition is the Cartesian structural condition corresponding to a decomposable class; it does not by itself determine whether the learner is proper or improper.

Assumption 1 (Bounded log ratios)**.** *For every relevant h and $q_h \in \mathcal{Q}_0$,*

$$\left| \log \frac{q^\star_h(U_h \mid U_{<h})}{q_h(U_h \mid U_{<h})} \right| \le G$$

almost surely under $P^\star$, for some finite $G > 0$.

Table 1: Observable-history continuation models in three representative domains.

Domain	Observable history	Intervention	Structured continuation and decision
Pricing	prices, sales, availability, calendar	price	demand/sales law; optimize revenue or risk
Inventory	orders, sales, stockouts, arrivals	order or allocation	lost-sales balance; optimize holding and shortage cost
Queueing	arrivals, departures, workload, abandonment	staffing or routing	Lindley/flow reactor; optimize capacity and delay

3.2 Linear statistical lower bounds

The first result shows that linear horizon dependence is unavoidable in a constructed realizable decomposable problem even for an arbitrary trajectory-law estimator. A second result establishes the same worst-case order for proper selection over a constructed realizable fully shared class.

Theorem 2 (Linear decomposable lower bound for arbitrary estimators)**.** *There exist universal constants $c, c_0 > 0$ such that the following holds. For every integer $M \geq 4$, every $n \geq c_0 \log M$, every horizon $H \in \mathbb{N}$, and every $G \in (0, 1]$, one can construct a single-step density class $\mathcal{Q}_0$ of cardinality M satisfying Assumption 1 such that, for the product family $\{P_q : q \in \mathcal{Q}_0^H\}$, the minimax joint-KL risk satisfies*

$$\inf_{\widehat{P}} \sup_{q^\star \in \mathcal{Q}_0^H} \mathbb{E}_{P_{q^\star}^{\otimes n}} \left[\mathrm{KL}\Big(P_{q^\star} \Big\| \widehat{P}(\mathcal{D}_n) \Big) \right] \geq c \, \frac{HG^2 \log M}{n}. \tag{10}$$

Here $\mathcal{D}_n$ denotes the n observed trajectories, and the infimum is over all possibly randomized measurable estimators that output an arbitrary probability law on the trajectory space. In particular, $\widehat{P}$ need not belong to the product family, factor coordinatewise, or select a candidate.

Thus properness is not essential for the decomposable lower bound: its factor H comes from H independently varying statistical coordinates. This strengthens the preliminary version, whose displayed Fano proof decoded a class member and therefore covered only proper selectors.

This is a lower bound for a constructed decomposable class, not for every shared-parameter model. It establishes that the H/n statistical term in the next theorem is not a consequence of a loose trajectory-level proof.

Correction to the preliminary version. The initial arXiv version of Xu et al. (2026) formally stated its lower bound for proper selectors, but its abstract, Table 1.1, contribution summary, and surrounding information-theoretic language did not consistently expose that restriction. It also asserted that the decomposable construction proved the fully shared case. That assertion does not follow: restricting only the selector to the shared diagonal while leaving the truth in the decomposable family yields a misspecified raw-risk comparison, not realizable shared minimax risk or excess risk relative to the best shared model. The detailed audit in Appendix A also corrects the preliminary proof's cardinality notation and mutual-information step and explains why its separate dependent-class claim is not used here. The next theorem supplies a genuinely shared construction, with its necessary properness qualifier.

Theorem 3 (Proper linear lower bound for a fully shared class)**.** *There exist universal constants $c, c_0 > 0$ such that the following holds. For every integer $M \geq 4$, every $n \geq c_0 \log M$, every horizon $H \in \mathbb{N}$, and every $G \in (0, 1]$, one can construct an observed-baseline context experiment and a fully shared class $\mathcal{Q}_{\mathrm{sh}} = \{q_1, \ldots, q_M\}$ satisfying Assumption 1. The same conditional kernel q_k is used at all H rounds, and, writing P_k for the resulting joint law of the context and length-H continuation,*

$$\inf_{\hat{q}} \sup_{k \in [M]} \mathbb{E}_{P_k^{\otimes n}} \big[\mathrm{KL}(P_k \| P_{\hat{q}}) \big] \geq c \, \frac{HG^2 \log M}{n}. \tag{11}$$

The infimum is over all possibly randomized, measurable $\mathcal{Q}_{\mathrm{sh}}$-valued estimators. Thus the result concerns proper model selection; it does not claim the same lower bound for unrestricted improper trajectory mixtures.

For G bounded away from zero, this lower bound matches the fully shared ERM upper bound below in H, n, and $\log M$. It is again a worst-case existence result: parameter sharing can provide substantial effective replication in other shared classes. Appendix C.5 gives the construction and also explains why merely tying the original product-family parameters cannot prove the result.

Proposition 2 (Improper aggregation of finite trajectory classes)**.** *Let $\mathcal{P} = \{P_1, \dots, P_N\}$ be any finite collection of trajectory laws. For every $n \geq 0$, there is an estimator $\widehat{P}_{\mathrm{mix}} \in \mathrm{conv}(\mathcal{P})$ based on n iid trajectories such that, for every trajectory law $P^\star$,*

$$\mathbb{E}_{(P^\star)^{\otimes n}} \mathrm{KL}(P^\star \| \widehat{P}_{\mathrm{mix}}) \leq \min_{k \in [N]} \mathrm{KL}(P^\star \| P_k) + \frac{\log N}{n+1}. \tag{12}$$

No horizon, boundedness, realizability, or autoregressive-independence assumption is required.

In the realizable decomposable family, $N = M^H$, so this upper bound is $H \log M/(n+1)$ and is consistent with Theorem 2. In a realizable fully shared family, $N = M$, so it is $\log M/(n+1)$, independently of H. Consequently, no $\Omega(HG^2 \log M/n)$ lower bound can hold uniformly over unrestricted improper estimators for finite fully shared classes. Appendix C.6 proves the proposition by progressive posterior aggregation.

3.3 The oracle upper bounds

Let $\hat{q}$ minimize empirical joint log loss over the candidate class:

$$\hat{q} \in \operatorname*{argmin}_{q \in \mathcal{Q}} -\frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{H} \log q_h(U_h^i \mid U_{<h}^i). \tag{13}$$

Theorem 4 (Autoregressive oracle inequality in joint KL)**.** *Suppose Assumption 1 holds and $\mathcal{Q}_0$ is finite. There is a universal constant C such that for every $\varepsilon \in (0, 1]$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$:*

1. *If $\mathcal{Q} = \mathcal{Q}_0^H$ is decomposable, then*

$$\mathrm{KL}(P^\star \| P_{\hat{q}}) \leq (1+\varepsilon) \inf_{q \in \mathcal{Q}_0^H} \mathrm{KL}(P^\star \| P_q) + C \frac{(1+\varepsilon)^2}{\varepsilon} \frac{H(G+1) \log(2H|\mathcal{Q}_0|/\delta)}{n}. \tag{14}$$

2. *If $\mathcal{Q} = \mathcal{Q}_{\mathrm{sh}}$ is fully shared, then*

$$\mathrm{KL}(P^\star \| P_{\hat{q}}) \leq (1+\varepsilon) \inf_{q \in \mathcal{Q}_{\mathrm{sh}}} \mathrm{KL}(P^\star \| P_q) + C \frac{(1+\varepsilon)^2}{\varepsilon} \frac{(HG+1) \log(2|\mathcal{Q}_0|/\delta)}{n}. \tag{15}$$

For decomposable classes, the proof concentrates over the coordinate-wise index set (h, q) and then sums conditional KL through (9); it does not treat the entire HG-bounded trajectory loss over the product class as one observation. This is what avoids a naive H^2 remainder. The following theorem extends the result to infinite model classes via log-ratio covers, with covering numbers defined under the uniform log-ratio distance in (7). This definition also corrects a notational typo in the initial version of Xu et al. (2026), where the absolute value was inadvertently omitted.

Table 2: Scope of the realizable finite-class joint-KL results, with $|\mathcal{Q}_0| = M$. "Proper" means class-valued; "improper" permits an arbitrary complete-trajectory law, following the learner-scope distinction in Rohatgi et al. (2025). Every lower bound is a worst-case construction, not a claim about every class of the displayed form.

Class structure	Proper class-valued selector	Unrestricted improper trajectory-law estimator
Decomposable: $\lvert\mathcal{Q}_0^H\rvert = M^H$	Constructed lower bound $\Omega(HG^2 \log M/n)$.	The same lower bound holds. Generic aggregation is at most $H \log M/(n+1)$, so the linear-H dependence is tight for $G = \Theta(1)$.
Fully shared: $\lvert\mathcal{Q}_{\mathrm{sh}}\rvert = M$	Constructed lower bound $\Omega(HG^2 \log M/n)$ from the rare-context experiment.	Generic aggregation is at most $\log M/(n+1)$; a uniform linear-H lower bound is therefore impossible.

Theorem 5 (Autoregressive oracle inequality with log-ratio covers)**.** *Use the uniform Radon–Nikodym log-ratio distance $d_{\log}$ defined in Section 2, with time included in the conditioning variable. Let $\mathcal{Q}_{0,r} \subseteq \mathcal{Q}_0$ be an r-net in this distance with $|\mathcal{Q}_{0,r}| = N_{\log}(\mathcal{Q}_0, r)$, and let $\hat{q}_r$ be an ERM over $\mathcal{Q}_{0,r}^H$. Replace finiteness of $\mathcal{Q}_0$ in Theorem 4 by the existence of the displayed finite r-net, and retain its bounded log-ratio condition. Then, with probability at least $1-\delta$,*

$$\mathrm{KL}(P^\star \| P_{\hat{q}_r}) \le (1+\varepsilon) \inf_{q \in \mathcal{Q}_0^H} \mathrm{KL}(P^\star \| P_q) + C \frac{(1+\varepsilon)^2}{\varepsilon} \frac{H(G+r+1)\log(2HN_{\log}(\mathcal{Q}_0, r)/\delta)}{n} + C' Hr. \tag{16}$$

Here C' may depend on ε only through a universal multiple of $1+\varepsilon$. For the fully shared class, use an ERM over the corresponding shared net, replace $\log(2HN_{\log}(\mathcal{Q}_0, r)/\delta)$ by $\log(2N_{\log}(\mathcal{Q}_0, r)/\delta)$, and replace $H(G+r+1)$ by $H(G+r)+1$.

3.4 The horizon message

Together, the results separate structure, learner scope, and approximation:

$$\begin{aligned} \text{decomposable, arbitrary output} &: \Omega(HG^2 \log M/n) \text{ can occur,} \\ \text{fully shared, proper output} &: \Omega(HG^2 \log M/n) \text{ can occur,} \\ \text{fully shared, improper output} &: \text{excess KL} \le \log M/(n+1), \\ \text{joint-KL approximation} &: \inf_{q \in \mathcal{Q}} \mathrm{KL}(P^\star \| P_q) \text{ gets no extra factor } H. \end{aligned}$$

The last line does not say that joint-KL approximation is numerically independent of the horizon; the joint divergence already contains the full trajectory. It says that the oracle inequality does not multiply that best-in-class trajectory error by another factor of H. This alignment is the central statistical advantage of working directly in joint KL rather than converting to a sequence metric with only an approximate chain rule. The lower-bound lines refer to constructed worst-case classes; Table 2 records their exact scope.

4 From Continuation Laws to Decisions

Statistical risk is first measured under the trajectory distribution that generates the data. Decision guarantees concern the distribution induced by a candidate intervention. The distinction is explicit in the following bridge.

Proposition 3 (Logging-to-decision KL transfer)**.** *Fix a learned continuation law q. Let $\rho^\star_\nu$ denote the true time-indexed history–action occupancy measure on the disjoint union $\bigcup_{t=1}^{H}\{t\}\times\mathcal{H}_t\times\mathcal{A}_t$ under a decision kernel ν. Assume the true and learned trajectory laws use the same initial distribution and the same decision kernel. Then*

$$\mathrm{KL}(P^\star_\nu\|P^q_\nu)=\sum_{t=1}^{H}\mathbb{E}_{P^\star_\nu}\mathrm{KL}(q^\star_t(\cdot\mid H_t,A_t)\|q_t(\cdot\mid H_t,A_t))\,. \tag{17}$$

In particular, let μ be the logging kernel and d a candidate decision kernel. If $\rho^\star_d\ll\rho^\star_\mu$ and $d\rho^\star_d/d\rho^\star_\mu\le C_{\mathrm{cov}}$, then

$$\mathrm{KL}(P^\star_d\|P^q_d)\le C_{\mathrm{cov}}\mathrm{KL}(P^\star_\mu\|P^q_\mu). \tag{18}$$

The proposition makes clear why randomized logging is useful and why no training guarantee can validate unsupported counterfactuals.

4.1 Generic, fast, and reusable decision transfer

Fix a history h and a class $D(h)$ of one-step actions, open-loop plans, or closed-loop continuation rules. For each $d\in D(h)$, let $P^\star_d$ and $P^{\hat q}_d$ be the true and learned continuation laws, and define

$$J^\star(d)=\mathbb{E}_{P^\star_d}[F_d],\qquad J_{\hat q}(d)=\mathbb{E}_{P^{\hat q}_d}[F_d],\qquad \Delta_D(h;\hat q)=\sup_{d\in D(h)}\mathrm{KL}(P^\star_d\|P^{\hat q}_d).$$

Let $d^\star$ and $\hat d$ maximize $J^\star$ and $J_{\hat q}$, respectively.

Theorem 6 (Decision transfer from a learned observable-history law)**.** *Assume the displayed maximizers exist.*

1. ***Generic bounded objectives.*** *If every F_d is measurable and has range at most B, then*

$$J^\star(d^\star)-J^\star(\hat d)\le B\sqrt{2\Delta_D(h;\hat q)}. \tag{19}$$

2. ***Worst-case sharpness.*** *For every sufficiently small $\kappa>0$ and $B>0$, there is a one-step problem with $\Delta_D(h;\hat q)\le\kappa$ and plug-in regret at least $cB\sqrt{\kappa}$ for a universal $c>0$.*

3. ***Fast transfer under stochastic-optimization stability.*** *Suppose $D(h)$ is convex, $J^\star$ and $J_{\hat q}$ are differentiable, $J^\star$ is μ-strongly concave (equivalently, $-J^\star$ is μ-strongly convex) with respect to a norm $\|\cdot\|$ whose dual is $\|\cdot\|_*$, and*

$$\sup_{d\in D(h)}\|\nabla J_{\hat q}(d)-\nabla J^\star(d)\|_*\le L_{\mathrm{KL}}\sqrt{\Delta_D(h;\hat q)}.$$

   Then

$$J^\star(d^\star)-J^\star(\hat d)\le\frac{L^2_{\mathrm{KL}}}{2\mu}\Delta_D(h;\hat q). \tag{20}$$

4. ***Objective reuse.*** *The induced laws do not depend on the downstream objective. Hence the same learned continuation law satisfies (19) for every new family of objectives with the corresponding range bound. Chance constraints and discontinuous risk criteria require their own slack, margin, or continuity conditions.*

Combining the proposition and theorem makes the role of coverage explicit. If $d\rho_d^\star/d\rho_\mu^\star \le C_{\mathrm{cov}}$ uniformly for $d \in D(h)$ for rollouts initialized at h, then

$$\Delta_D(h;\hat{q}) \le C_{\mathrm{cov}}\mathrm{KL}(P_\mu^\star \| P_\mu^{\hat{q}}).$$

Consequently, the generic regret bound becomes $B\sqrt{2C_{\mathrm{cov}}\mathrm{KL}(P_\mu^\star \| P_\mu^{\hat{q}})}$, and the fast bound becomes $(L_{\mathrm{KL}}^2 C_{\mathrm{cov}}/(2\mu))\mathrm{KL}(P_\mu^\star \| P_\mu^{\hat{q}})$.

Remark 1 (Where strong convexity lives)**.** *The curvature assumption in part 3 is imposed on the* downstream decision variable *d: reward is strongly concave in d, or equivalently loss is strongly convex in d. It is not an assumption that the autoregressive negative log likelihood is strongly convex in neural-network weights. This distinction matters. GRU and Transformer likelihoods are generally nonconvex and have parameter symmetries, so global strong convexity in their raw weights is normally unavailable; none is needed for* (20) *once a learned law and the stated prediction-to-gradient stability are given.*

In a realizable regular parametric model $q = q_\theta$, one useful local *route to that stability is*

$$\sup_d \|\nabla J_\theta(d) - \nabla J_{\theta^\star}(d)\|_* \le L_\theta \|\theta - \theta^\star\|, \qquad \Delta_D(h;q_\theta) \ge c_{\mathrm{KL}}\|\theta - \theta^\star\|^2,$$

throughout an identifiable neighborhood of $\theta^\star$. Then $L_{\mathrm{KL}} = L_\theta/\sqrt{c_{\mathrm{KL}}}$. The second inequality is local quadratic identifiability of the induced trajectory law, not a requirement of global strong convexity of the training problem. It follows, for example, for regular conditional exponential-family or generalized-linear autoregressive models when the relevant trajectory Fisher information is uniformly positive definite. For an overparameterized neural model it generally fails in raw parameter norm unless symmetries are quotiented out or the analysis is restricted to an identifiable subspace; a prediction-space KL condition is then the cleaner formulation. Adding an ℓ_2 penalty may make a convex linear fitting problem strongly convex, but by itself it establishes neither trajectory-KL identifiability nor downstream decision curvature.

If d itself parameterizes an autoregressive policy, strong concavity in d is likewise not automatic and usually fails for unrestricted neural policy weights. It is plausible in structured convex formulations—for example, continuous pricing with a strictly concave revenue curve, open-loop controls with linear dynamics and strongly convex costs, or suitably regularized optimization over action distributions or occupancy measures. Outside such settings one should use the generic square-root-KL guarantee, or prove a local margin, quadratic-growth, or optimizer-stability condition directly. Thus continuation learning preserves the probabilistic object through which operational curvature can be used; it does not make every decision problem fast.

4.2 A likelihood–value/policy approximation separation

The generic square-root rate can carry a horizon-sized objective range. The next construction has one time-zero intervention followed by H payoff periods; equivalently, it is an $(H+1)$-stage finite-horizon problem. Values are indexed by time, and sup norms are over state–time pairs.

Theorem 7 (Horizon-sensitive likelihood–decision separation)**.** *Fix $H \in \mathbb{N}$ and $\varepsilon \in (0, 1/4)$. There exist a finite decision problem, a true continuation law $q^\star$, and a misspecified singleton class $\mathcal{Q} = \{\bar{q}\}$ such that*

$$\kappa_\varepsilon := \sup_\pi \mathrm{KL}(P_{q^\star}^\pi \| P_{\bar{q}}^\pi) = -\frac{1}{2}\log(1 - 4\varepsilon^2) = \Theta(\varepsilon^2).$$

Nevertheless, the policy optimal under $\bar{q}$ has true regret at least $H\varepsilon/2$. Moreover, with time-indexed optimal values,

$$\|V^\star_{\bar{q}} - V^\star_{q^\star}\|_\infty \geq \frac{H\varepsilon}{2}, \qquad \|Q^\star_{\bar{q}} - Q^\star_{q^\star}\|_\infty \geq H\varepsilon,$$

where the lower bounds are witnessed at the time-zero state. Thus transferred policy and value errors can be $\Omega(H\sqrt{\kappa_\varepsilon})$.

What the theorem separates. Because $\mathcal{Q} = \{\bar{q}\}$ is a singleton, no estimation is involved: κ_ε is exactly the best-in-class worst-policy joint-KL approximation error for the full observable history. Yet exact selection of this oracle model induces policy regret and $V^\star$- and $Q^\star$-approximation errors at least of order $H\sqrt{\kappa_\varepsilon}$. Even after dividing returns and values by H, the exponent separation remains: trajectory KL is $\Theta(\varepsilon^2)$, whereas the normalized policy and value errors are $\Omega(\varepsilon) = \Omega(\sqrt{\kappa_\varepsilon})$. For example, for fixed $c > 0$ and all sufficiently large H, taking $\varepsilon_H = c/H$ makes the joint KL $\Theta(H^{-2})$ while the unnormalized policy and value errors remain bounded away from zero.

This is a model-to-policy and model-to-value sensitivity result, not repeated accumulation of independent Bellman residuals. It saturates the generic bound because the trajectory reward range is $B = H$. It is also a minimal benchmark for realizability sensitivity, not a lower bound for every robust RL algorithm or for independently chosen direct policy and value classes. The strongly concave pricing construction in Section 5.5 identifies the favorable geometric boundary: KL-stable gradients improve decision regret from $\Theta(H\sqrt{\kappa})$ to $\Theta(H\kappa)$, although the coefficient of the total-return guarantee can still scale with H.

5 Experimental Evidence

The experiments test the representation, statistical, and decision-transfer claims separately. They are not a broad forecasting leaderboard. Complete data-generating processes, estimands, hyperparameters, inference procedures, software versions, hashes, and audit gates appear in Appendix D.

5.1 Design and claim boundaries

The structural model learns a categorical shock law $p_\theta(\xi_{t+1} \mid h_t, a_t)$ and passes it through the known reactor. If an observation has multiple latent preimages, training uses the exact observed-data likelihood

$$-\log \sum_{\xi:\, \Phi_t(h_t, a_t, \xi) = z} p_\theta(\xi \mid h_t, a_t). \tag{21}$$

The primary encoder is a one-layer GRU; a lag-feature MLP diagnoses the value of recurrent history. The matched direct GRU predicts a global categorical law over observable continuations and is evaluated both raw and after prespecified support projection.

Synthetic systems use 900 independent 30-period episodes per seed, split by whole episode into 60% training, 20% validation, and 20% testing. Five complete data-generation/training seeds form the uncertainty units. Rossmann uses three neural initializations and a chronological split. All hyperparameters and stopping epochs use validation observed-continuation NLL only; all 39 neural fits stop before their epoch caps. Synthetic intervals are paired Student-t intervals over seeds. Rossmann's primary intervals describe cross-store variation conditional on the three-seed ensemble, with seed-level and two-way seed/store sensitivities reported separately.

Synthetic policies are scored under known true laws. Rossmann costs are one-step counterfactual evaluations on held-out states generated by the synthetic logging system. No evaluated action

changes a later state. Accordingly, the Rossmann comparison is neither a recursive deployment nor a causal field-savings estimate.

5.2 Representation: lost-sales inventory and a reflected queue

Inventory obeys

$$Y_t = I_t + a_t, \qquad S_{t+1} = \min\{D_{t+1}, Y_t\}, \qquad I_{t+1} = Y_t - S_{t+1},$$

with censored demand when $S_{t+1} = Y_t$. The queue obeys the controlled Lindley recursion (Lindley, 1952), $Q_{t+1} = [Q_t + X_{t+1}]^+$; at zero, several negative net-input shocks produce the same observation. Both logging rules use 0.35 randomized exploration. Orders minimize $0.35I_{t+1} + 21\{I_{t+1} = 0\}$, and staffing minimizes $0.30a_t + \mathbb{E}Q_{t+1}$.

Table 3 reports the held-out representation and one-step decision comparisons.

Table 3: Synthetic structural-model results. Entries are means with 95% Student-t intervals over five independent complete runs. Lower is better. Boldface marks the smallest point estimate among compared model variants within each domain and metric; ties are all marked. It is descriptive and does not by itself indicate a statistically resolved difference. The true-law oracle used to define regret is excluded.

Domain	Model	Observable KL	Invalid probability	Decision regret
Inventory	Structural GRU	**0.00446 (0.00382, 0.00510)**	**0**	**0.00263 (0.00185, 0.00341)**
	Structural MLP	0.01627 (0.01322, 0.01932)	**0**	0.00810 (0.00596, 0.01025)
	Direct GRU	0.04180 (0.03158, 0.05202)	0.01802 (0.01135, 0.02470)	0.01196 (0.00880, 0.01511)
	Direct GRU + projection	0.02126 (0.01839, 0.02412)	**0**	0.01210 (0.00919, 0.01501)
Queue	Structural GRU	**0.00357 (0.00328, 0.00385)**	**0**	**0.00103 (0.00061, 0.00144)**
	Structural MLP	0.01729 (0.01646, 0.01813)	**0**	0.01362 (0.00827, 0.01897)
	Direct GRU	0.00780 (0.00700, 0.00860)	0.00169 (0.00133, 0.00204)	0.00116 (0.00060, 0.00173)
	Direct GRU + projection	0.00609 (0.00544, 0.00675)	**0**	0.00150 (0.00048, 0.00253)

The structural GRU assigns exactly zero infeasible probability in both systems. In inventory, its observable KL is 0.00446, compared with 0.02126 after direct support projection; the paired reduction is 0.01680 with 95% interval $[0.01395, 0.01965]$. Its regret is also lower than both direct variants with paired intervals separated from zero. In the queue, observable KL is 0.00357, compared with 0.00609 after projection; the paired reduction is 0.00253 with interval $[0.00175, 0.00331]$. Queue-regret point estimates favor the structural GRU, but the paired differences from the direct variants are not resolved.

Figure 2 displays feasibility, data processing, and decision loss for the same two systems.

Exact enumeration gives maximum data-processing excess 8.33×10^{-17}, numerical roundoff. The independent-trajectory chain-rule audit is within 1.92 Monte Carlo standard errors in every primary structural-GRU seed; Appendix Figure 5 reports the complete curves.

5.3 Observable histories under realistic serial variation

Public Rossmann Store Sales paths (Kaggle, 2015) introduce serial and calendar variation into the lost-sales reactor. Aggregate sales are converted into a training-scaled demand proxy; inventory, randomized orders, censoring, and costs are synthesized. The first 40 store identifiers with complete training coverage are selected without using validation or test missingness. Training, validation, and testing are chronological, and store scaling uses training dates only. Contemporaneous customer counts are excluded, and the learner never observes the uncensored demand proxy after stockout.

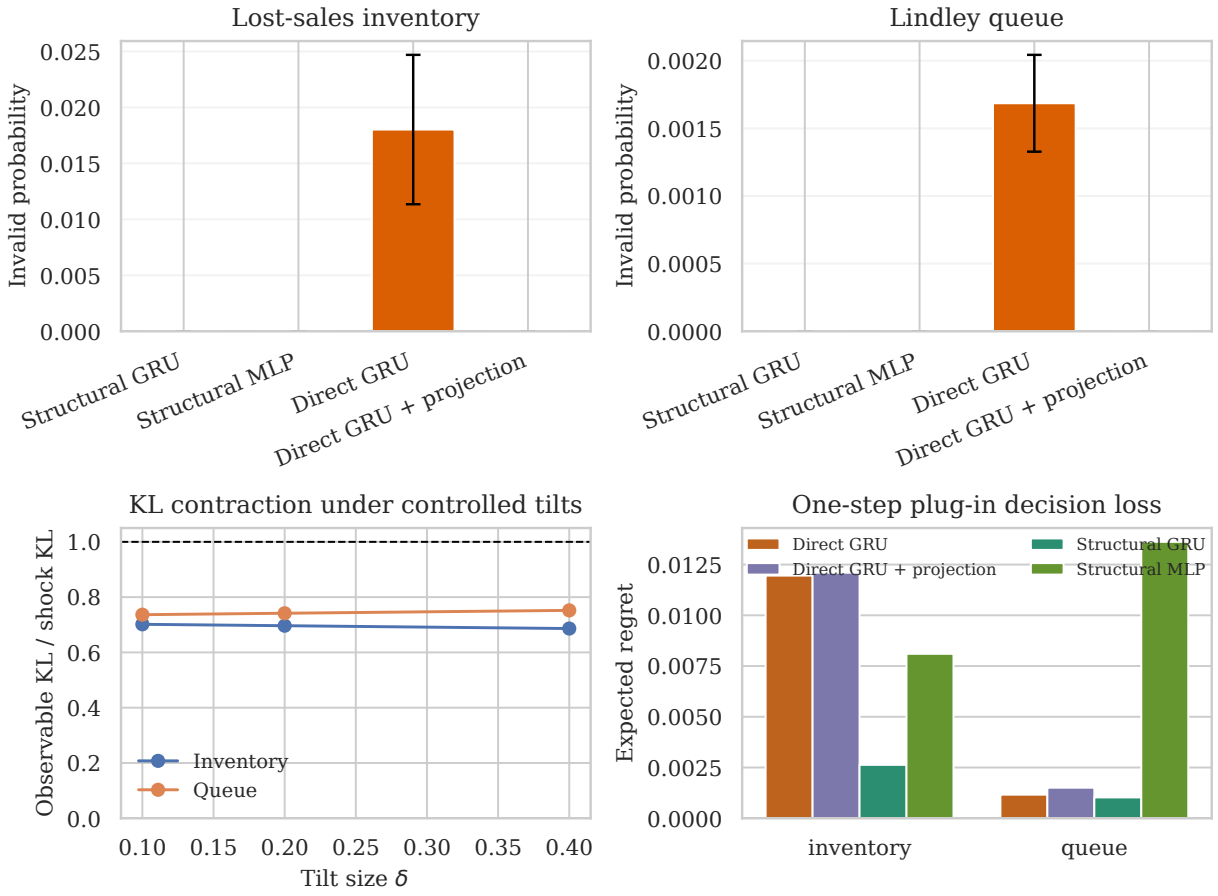

Figure 2: Representation effects in lost-sales inventory and a reflected queue. The panels report invalid mass, exact KL contraction under controlled tilts, and one-step plug-in regret. Neural intervals use five independent synthetic seeds.

Table 4: Semi-synthetic Rossmann results. Neural initializations are averaged within store; entries are means with conditional 95% intervals over 40 store clusters. Lower is better. Boldface marks the smallest point estimate among compared model variants in each column; ties are all marked. It is descriptive and does not by itself indicate a statistically resolved difference. Logging and perfect-information reference policies are excluded from the ranking.

Model	Observed NLL	Invalid probability	Constructed cost/day
Structural GRU	**0.3017 (0.2415, 0.3619)**	**0**	**1.0758 (1.0107, 1.1409)**
Structural MLP	0.3130 (0.2529, 0.3731)	**0**	1.0812 (1.0205, 1.1419)
Direct GRU	0.3628 (0.2807, 0.4449)	0.02963 (0.02440, 0.03485)	1.0867 (1.0225, 1.1508)
Direct GRU + projection	0.3231 (0.2489, 0.3972)	**0**	1.0896 (1.0265, 1.1527)

Table 4 reports the held-out semi-synthetic comparison.

The structural GRU lowers observed NLL by 0.0611 relative to the raw direct GRU and assigns no infeasible mass. Its constructed one-step cost is 1.0758 per store-day, compared with 1.6191 for the randomized logging rule and a perfect-information lower bound of 0.9221. The 0.5433 reduction is 33.6% of logging cost; its seed-level interval is $[0.5354, 0.5512]$ and its two-way seed/store sensitivity interval is $[0.4677, 0.6190]$. Smaller cost differences among learned models are directional and can include zero under two-way inference. Appendix Figure 6 supplies the graphical diagnostic. These results concern the synthesized inventory objective, not Rossmann field operations.

5.4 Statistical accumulation in joint KL

At each coordinate, the true Bernoulli parameter belongs to

$$\left\{\tfrac{1}{2} - 3\delta_n, \tfrac{1}{2} - \delta_n, \tfrac{1}{2} + \delta_n, \tfrac{1}{2} + 3\delta_n\right\}, \qquad \delta_n = \frac{1}{2\sqrt{n}},$$

and an independent maximum-likelihood selector is fit. We use $H \in \{5, 10, 20, 40, 80\}$, $n \in \{100, 250, 500, 1000, 2000\}$, and 1,000 replications per cell. Across all 25 cells,

$$0.4715 \leq \frac{n\,\mathbb{E}[\mathrm{KL}]}{H} \leq 0.5035.$$

The experiment recovers the local worst-case H/n mechanism; because the local class changes with n, it is not a learning-curve claim for a fixed benign GRU.

Figure 3 shows both the raw accumulation and its n/H normalization.

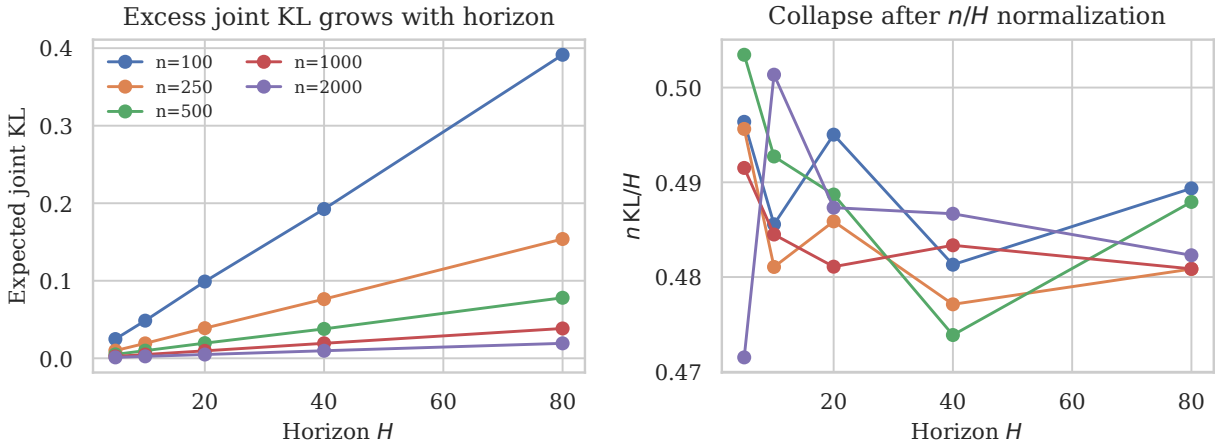

Figure 3: Finite-class joint-KL scaling. Expected joint KL is linear in H and inverse in n; the normalized quantity $n\mathrm{KL}/H$ is nearly constant over the full grid.

5.5 Decision geometry under a common misspecification

This is the computational counterpart to Theorem 7. It uses exactly the same pair of true and misspecified continuation laws in two decision problems, thereby holding the H-independent law-approximation error fixed and changing only the geometry of decision extraction. At time zero a manager commits to one price for H periods. A persistent market state satisfies $Y \sim \mathrm{Bern}(1/2)$

under truth and $Y \sim \mathrm{Bern}(1/2+\delta)$ under the misspecified model, with $D_t(p) = 4 - p + Y$ and profit $pD_t(p)$. Because Y is drawn only once, the full observable-trajectory KL is

$$\kappa_\delta = -\tfrac{1}{2}\log(1-4\delta^2) = \Theta(\delta^2),$$

independently of H.

On the continuous convex price set $[1,3]$, the true and model expected total profits are

$$J_H^\star(p) = Hp(4.5-p), \qquad \bar{J}_H(p) = Hp(4.5+\delta-p).$$

Both are $2H$-strongly concave in the *decision* p—equivalently, negative total profit is $2H$-strongly convex. Their optimizers are $p^\star = 2.25$ and $\hat{p} = 2.25 + \delta/2$, and the exact true regret is $H\delta^2/4 = \Theta(H\kappa_\delta)$. Moreover,

$$\sup_{p\in[1,3]} \left|\bar{J}_H'(p) - J_H^{\star\prime}(p)\right| = H\delta \le H\sqrt{\kappa_\delta/2}, \qquad \mu = 2H.$$

Part 3 of Theorem 6 therefore gives the explicit upper bound $H\kappa_\delta/8$, which is locally attained to first order because $\kappa_\delta \sim 2\delta^2$. This is the linear-KL side of the separation.

For the prespecified two-price near-tie, the continuation laws and ambient quadratic profits are unchanged, but the feasible set contains only two points. The model switches to the wrong price and incurs exact regret $H\delta/2 = \Theta(H\sqrt{\kappa_\delta})$. Although the ambient quadratic is curved, the two-point feasible set is nonconvex, the convex-domain first-order condition in the fast-transfer theorem is unavailable, and the argmax switches discontinuously at the tie. Thus curvature of an ambient formula alone is not enough: convex decision geometry and optimizer stability change the misspecification exponent from $1/2$ to 1, while the scale of unnormalized total profit remains linear in H.

The autoregressive model in this diagnostic is a one-parameter Bernoulli law, but the fast rate is not derived from strong convexity of its likelihood in that parameter. It follows from strong concavity in price plus the displayed KL-to-profit-gradient inequality. The same conclusion would hold if a nonconvex neural autoregressive parameterization induced these two continuation laws; see Remark 1.

Across the small-δ grid, fitted log–log slopes of per-period regret against joint KL are 0.999 for the strongly concave problem and 0.499 for the discrete near-tie, matching the theoretical exponents 1 and $1/2$.

Figure 4 verifies both transfer exponents and their horizon scaling.

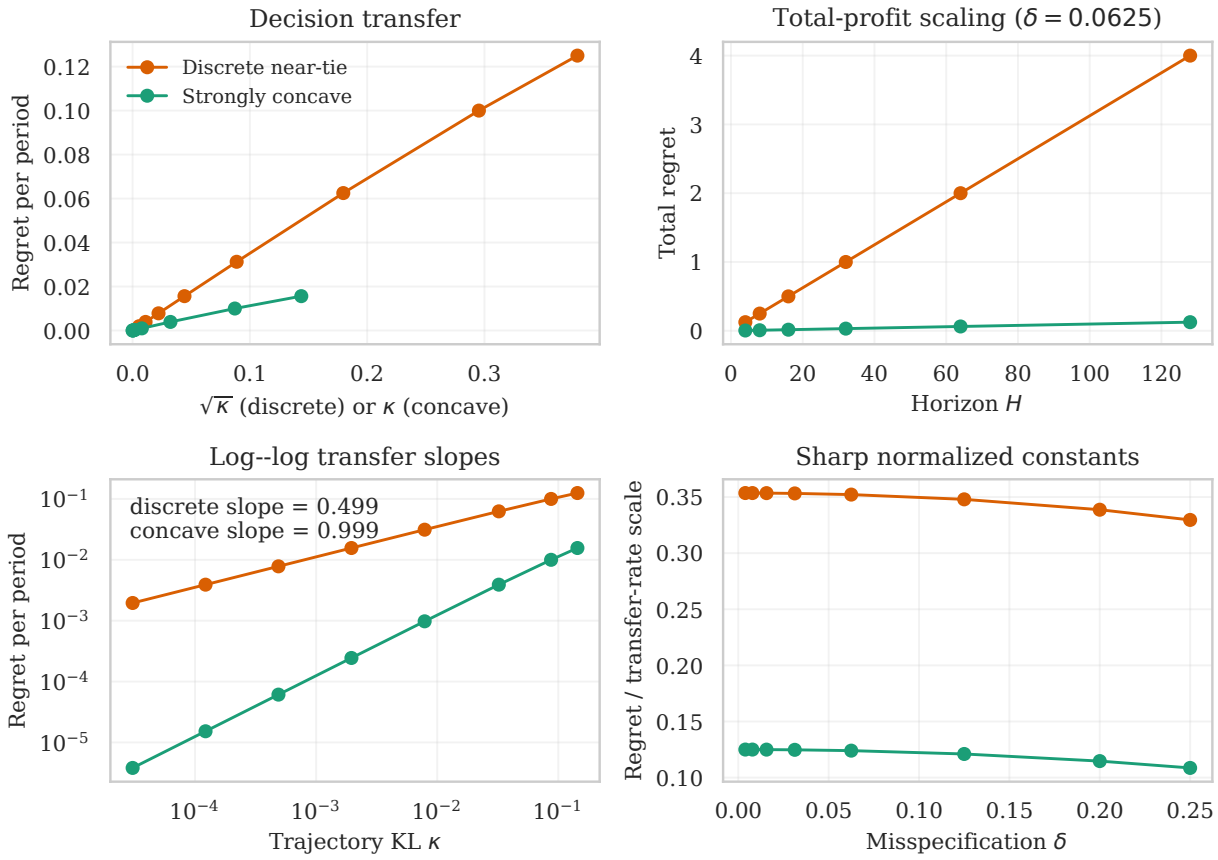

Figure 4: Decision geometry under the same H-independent trajectory misspecification. A discrete near-tie produces square-root-KL sensitivity, whereas strong concavity in the continuous price decision, together with KL-stable gradients, stabilizes the optimizer and yields linear-KL regret.

Taken together, the experiments support four bounded claims: observable reactors enforce feasibility; they contract KL at censoring and reflection boundaries; the constructed finite-class diagnostics exhibit a linear statistical horizon cost under the learner scopes in Table 2; and decision geometry, not predictive loss alone, determines the final transfer rate. They do not establish universal dominance of structural models or the performance of a recursively deployed Rossmann policy.

6 Conclusion

Autoregressive decision making should begin with a choice of what to learn. When actions are interventions and the system is observed through rich, partially observed histories, the controlled observable continuation law is a natural primitive. It avoids an unverified Markov compression, admits exact operational reactors, and remains reusable when the objective changes.

The sources of these gains are distinct. Full history removes compression bias on covered actions; a correct reactor supplies feasibility and data processing; joint likelihood supplies the statistical theory; and downstream optimization geometry determines whether KL transfers at a square-root or linear rate.

The joint-KL theory gives this modeling choice a sharp statistical backbone. Separate realizable constructions show that linear horizon dependence is unavoidable in the worst case for arbitrary trajectory-law estimators over a decomposable finite class and for proper selectors over a genuinely fully shared finite class. The distinction is essential: unrestricted aggregation over M shared trajectory laws has excess KL at most $\log M/(n+1)$ and therefore cannot obey the same linear-H lower bound. Log-loss ERM attains the relevant H/n order up to logarithmic and boundedness factors. At the same time, the exact chain rule prevents an additional horizon multiplier on best-in-class joint-KL approximation. Downstream, generic bounded objectives have tight square-root sensitivity, while curved and stable stochastic programs can retain a linear-KL rate.

Horizon-free joint-KL approximation does not mean horizon-free approximation of every downstream policy or value objective. The singleton construction in Theorem 7 yields $H\sqrt{\kappa}$ induced policy, value, and action-value errors, whereas strong concavity of the downstream reward together with KL-stable gradients improves decision regret to the $H\kappa$ scale in persistent pricing. Realizability sensitivity is therefore jointly a property of the learned object, its approximation metric, objective normalization, and the geometry of decision extraction.

The framework is deliberately conditional on identification, coverage, and the correctness of any encoded reactor. These are not technical footnotes: they separate reliable consequence modeling from unsupported counterfactual claims. Within those boundaries, learning from observable histories offers a clean division of labor—probability models the consequences, operational structure constrains them, and optimization chooses the action.

Acknowledgments

The authors used AI tools to support research, coding, and editorial work. ChatGPT and Codex assisted with the implementation and auditing of the experiments through an approximately four-hour, multistage workflow, including a separate audit pass intended to identify potential errors. The authors independently designed the complete experimental pipeline to validate the paper's theoretical mechanisms. They reviewed the resulting outputs and are completing a comprehensive manual review of the code and experimental details. The authors retain full responsibility for the experimental design, implementation, results, and interpretation.

A Related Work and Additional Comparisons

A.1 Related work

Relation to the preliminary version. A preliminary short paper (Xu et al., 2026) develops the joint-KL oracle and minimax learning results, together with a basic Pinsker rollout implication, in an autoregressive next-token formulation. It does not formulate controlled observable-history decision making. The present paper adds that formulation, the history-compression identity, structural-reactor and coverage results, the objective-dependent joint-KL versus policy/value separation, and the experimental study.

The initial arXiv version had several distinct lower-bound limitations, which we record explicitly.

1. Its Theorem 4.2 ranges over selectors $\widehat{\pi} \in \Pi$, and its proof uses the requirement that a proper learner choose one packed policy. Thus it proved only a proper-selection statement. This restriction was visible in the formal theorem and appendix but not consistently in the abstract, Table 1.1, main-contribution summary, or surrounding description of an information-theoretic barrier.

2. The proof constructs a base class Π_0 and the product $\Pi = \Pi_0^H$, obtaining $HG^2 \log |\Pi_0|/n$. If $|\Pi|$ denotes the cardinality of the full product, this equals $G^2 \log |\Pi|/n$, not $HG^2 \log |\Pi|/n$. We use $M = |\Pi_0|$ throughout to remove this ambiguity.

3. The decomposable packing conclusion is recoverable, but not literally by the mutual-information step as written. The appendix asserted that the uniform mixture of nonzero pairwise-orthogonal exponential tilts was uniform; that would require their sign vectors to sum to zero, which pairwise orthogonality precludes. The proof of Theorem 2 instead uses a balanced Varshamov–Gilbert packing, pairwise-KL information control, and a new posterior-predictive argument that extends the conclusion to arbitrary output laws.

4. The claimed fully shared consequence restricted only the learner to a shared diagonal while retaining decomposable truths. This is misspecified raw risk, not a realizable shared minimax lower bound or excess risk relative to the best shared model. The fully shared entry in Table 1.1 was therefore not supported. Theorem 3 supplies a separate proper shared construction, and Proposition 2 proves that its properness qualifier cannot be removed.

5. The separate H^2 dependent-class corollary is also not established by its displayed proof. Substituting its chosen perturbation into the preceding mutual-information bound leaves an H^2G^2 term in Fano's numerator, so the claimed constant testing error does not follow uniformly; the construction also bounds a joint HG log ratio without exhibiting an H-step factorization satisfying the stated per-round G condition. The present paper does not use that claim.

Finally, the preliminary section called "improper learning" considers the particular enlargement from a coupled class to the Cartesian product of its coordinate projections. It is not unrestricted aggregation of complete trajectory laws. Table 2 separates these notions.

Data-driven optimization and contextual stochastic optimization. A large OR literature studies how data should be used to prescribe decisions under uncertainty. Bertsimas and Kallus (2020) connect predictive and prescriptive analytics by using covariates to inform decisions in stochastic optimization. Predict-then-optimize and decision-focused learning ask how predictive errors propagate through optimization, and in some cases how the optimization problem should

shape the learning objective (Donti et al., 2017; Elmachtoub and Grigas, 2022). Our framework belongs to this tradition: the ultimate output is a decision obtained by optimizing a criterion. The new feature is that the predictive object is not necessarily a scalar objective-value predictor or a parameter vector; it is a controlled autoregressive law over observable continuations.

Deep learning for operational decisions. Lagzi et al. (2026) propose using neural networks to guide data-driven stochastic optimization by predicting the objective value as a function of covariates and decisions, then optimizing the trained network. They combine data-driven optimization with flexible function approximation. We instead combine data-driven optimization with conditional generative modeling. We replace objective-value prediction by continuation-law prediction. This is not a cosmetic change: a conditional law can be reused across objectives, constraints, risk measures, and rollout plans, whereas a scalar objective predictor is tied to a particular objective.

Probabilistic forecasting and structure-aware prediction. Probabilistic forecasting models such as DeepAR learn autoregressive conditional distributions and have demonstrated the value of sharing statistical strength across related time series (Salinas et al., 2020). Our object differs in two ways. First, the candidate decision enters the conditional law as an intervention, so action support and identification are explicit. Second, a known inventory, queueing, or capacity reactor is retained outside the neural decoder. This makes feasibility a property of the model class and permits an exact observed-data likelihood under censoring or reflection. The experimental study compares this structural factorization with a direct categorical decoder that is expressive enough to represent atoms but is not given the context-specific feasible support.

World models, sequence models, and action-token methods. World-model approaches learn predictive models of future observations for planning or control (Ha and Schmidhuber, 2018). Decision Transformer casts reinforcement learning as conditional sequence modeling and generates future actions conditioned on desired returns and past trajectories (Chen et al., 2021). Recent OMGPT-style work formulates several OR/OM tasks as sequence modeling problems and emphasizes predicting future actions from histories (Wang et al., 2025). These works demonstrate the power of sequence models for decision tasks. Our distinction is mathematical: we do not learn $p(a_t \mid h_t)$ as the primary object. We learn $q^\star(z_{t+1} \mid h_t, a_t)$. The action is an intervention and the prediction target is the operational consequence.

Autoregressive learning, next-token prediction, and in-context learning. Autoregressive models factorize a joint law into conditional laws. Motivated by the maximum-likelihood objective used in large-scale pretraining (Radford et al., 2018), early theoretical studies developed Bayesian or latent-variable interpretations of in-context learning (Xie et al., 2022; Wies et al., 2023; Wang et al., 2023; Jiang, 2023). These works explain how autoregressive models can perform implicit inference from context, but they do not provide a misspecified joint-KL oracle theory for autoregressive density estimation. Closest to our perspective, Zhang et al. (2023) study next-token prediction under misspecification using joint, rather than trajectory-conditioned, evaluation metrics; their main guarantee is stated in joint total variation distance. Joint-KL-type quantities appear in their proofs and in-context-learning consequences, but their results do not yield a sharp estimation–approximation decomposition for joint KL. Relatedly, Yuksel and Flammarion (2025) obtain KL-type guarantees for log-loss training of autoregressive processes, but their approximation term is measured by a worst-case total-variation criterion rather than by the joint KL metric used for evaluation. Our contribution instead gives joint-KL upper bounds for both structures, an

arbitrary-output lower bound for a constructed decomposable class, and a proper-selection lower bound for a separately constructed fully shared class.

Hellinger analyses, learner scope, and parameter sharing. Closely related to the preceding line of work, a series of papers studies autoregressive and imitation learning under squared Hellinger distance. Foster et al. (2024) analyze log-loss training for general behavior cloning and obtain finite-sample guarantees that imply multistep control under Hellinger-type evaluation, with autoregressive learning as a special case. Rohatgi et al. (2025) establish comprehensive upper and lower bounds that are highly informative, and distinguish their contribution from earlier work on autoregressive learning by providing an oracle theory.

Importantly, their results also clarify how horizon dependence varies with learner scope and model-class structure. In particular, Theorem 4.7 identifies a loss-uniform next-token barrier for proper iterative learners over classes with no parameter sharing. Proposition F.8 shows that a broad family of additive layerwise objectives fits this abstraction under the Cartesian structure $\Pi = \Pi_1 \times \cdots \times \Pi_H$. This regime most closely parallels our decomposable classes under proper learning. Their Section 5 complements this analysis by studying computational hardness for autoregressive linear models with a parameter shared across rounds, a structure closer to our fully shared classes. Taken together, these results illuminate several important regimes defined by learner scope and parameter sharing, while the statistical approximation behavior of fully shared classes does not follow directly from them.

Accordingly, we refine the interpretation of Table 1.1 in the initial version of Xu et al. (2026): an $\Theta(H)$ Hellinger entry is not a metric-only minimax statement. Horizon amplification depends jointly on the divergence, the learning objective, the output restriction, and class structure. Exact KL chain rules prevent an extra approximation multiplier for the log-loss ERM analyzed here; specified next-token procedures under Hellinger evaluation can behave differently.

Dynamic learnability and stochastic-process structure. Recent learning-theoretic programs ask when dynamical systems are learnable from observations alone, emphasizing stability, mixing, observability, and other structural properties (Hazan et al., 2025). Our framework studies a controlled, decision-oriented analogue: when observable continuation laws under candidate interventions can be learned in joint KL well enough to support stochastic optimization. In operational systems, structure may come from balance equations, regeneration, finite memory, renewal cycles, or queueing and inventory physics. These structures are part of the operational grammar that makes the continuation law learnable.

A.2 Model misspecification, value transfer, and realizability

The horizon-sensitive construction is related to simulation-lemma analyses and error propagation in approximate dynamic programming (Lobel and Parr, 2024; Farahmand et al., 2010). Its object and metric are different: we perturb a controlled continuation law over observable histories, measure the resulting trajectory discrepancy in joint KL, and study the policy and value objects obtained after optimization. The construction contains one consequential intervention followed by persistent payoffs; it demonstrates $H\sqrt{\kappa}$ sensitivity, not repeated accumulation of independent Bellman residuals. This distinction is why the main text uses the more precise term "horizon-sensitive transfer."

Approximate realizability is therefore objective-dependent. The Simons Institute workshop frames robustness to realizability as a central problem for RL theory (Simons Institute, 2020).

Our theorem isolates one obstruction: a continuation class may be close in worst-policy trajectory KL yet far in the policy and value objects induced by optimizing that class. It does not address exploration and does not prove an impossibility for every direct policy or value class. The strongly concave companion supplies the favorable side: together with KL-stable gradients, decision geometry improves the misspecification exponent from 1/2 to 1, even when the total-objective coefficient retains its horizon scale.

A.3 Dependent model classes and stepwise lifting

When a feasible conditional class $\mathcal{Q} \subseteq \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_H$ couples coordinates, a proper trajectory-level learner may face a global selection problem. If $\mathcal{Q}_h$ denotes the hth coordinate projection and $\mathcal{Q}_0 = \bigcup_h \mathcal{Q}_h$, lifting to $\mathcal{Q}_0^H$ permits stepwise estimation and the decomposable oracle bound. This changes the model class and can change approximation error or structural feasibility; it is a relaxation, not a free computational result. In particular, calling this lift "improper" means only improper relative to the original coupled class. It is not the unrestricted convex aggregation of complete trajectory laws in Proposition 2, and a lower bound for the lifted procedure is not automatically a lower bound for every improper learner.

B Notation Summary

Symbol	Meaning
z_t	Observable operational information at time t
a_t	Operational decision/intervention at time t
h_t	Observable history $(z_1, a_1, \ldots, z_t)$
$q^\star(z_{t+1} \mid h_t, a_t)$	True controlled observable continuation law
$\hat{q}$	Learned continuation law
P_q	Joint autoregressive trajectory law induced by conditionals $q_1, \ldots, q_H$
$K_h(q_h)$	Stepwise population conditional KL at horizon h
P_d^q	Joint law induced by model q and decision object d
$J_q(h, a)$	Conditional expected objective under model q

C Proofs of the Theoretical Results

C.1 Proof of Proposition 1

Proof. Because $S = \phi(H)$, conditioning on (H, S, A) is equivalent to conditioning on (H, A) for the law of Z. Insert the conditional law $P_{Z|S,A}$ into the log likelihood ratio:

$$\log \frac{dP_{Z|H,A}}{dr_{Z|S,A}}(Z) = \log \frac{dP_{Z|H,A}}{dP_{Z|S,A}}(Z) + \log \frac{dP_{Z|S,A}}{dr_{Z|S,A}}(Z).$$

Taking expectations, the first term is the conditional mutual information $I(Z; H \mid S, A)$, and the tower property turns the second into $\mathbb{E}\mathrm{KL}(P_{Z|S,A} \| r_{Z|S,A})$. This proves (5). Nonnegativity of KL shows that the minimum is achieved by $r = P_{Z|S,A}$ and equals the first term. Finally, $I(Z; H \mid S, A) = 0$ if and only if the corresponding regular conditional laws agree almost surely, equivalently $Z \perp H \mid (S, A)$. □

C.2 Proof of Theorem 1

Proof. Fix t, h, a, and abbreviate $T(\xi) = \Phi_t(h, a, \xi)$. For every $p_t \in \mathcal{P}_t$,

$$q_t^p(\mathcal{S}_t(h,a) \mid h, a) = p_t\big(T^{-1}(\mathcal{S}_t(h,a)) \mid h, a\big) = 1,$$

because $T(\xi) \in \mathcal{S}_t(h, a)$ for every ξ. This proves structural feasibility.

The data-processing inequality for relative entropy states that for every measurable map T,

$$\mathrm{KL}(T_{\#}P \| T_{\#}Q) \le \mathrm{KL}(P\|Q),$$

with the inequality understood in the extended real line. Applying it conditionally to $p_t^\star(\cdot \mid h, a)$ and $p_t(\cdot \mid h, a)$ gives

$$\mathrm{KL}(q_t^\star(\cdot \mid h,a) \| q_t^p(\cdot \mid h,a)) \le \mathrm{KL}(p_t^\star(\cdot \mid h,a) \| p_t(\cdot \mid h,a))\,.$$

Because $P_d^\star$ and $P_d^{q^p}$ use the same initial and decision kernels, those factors cancel in the trajectory likelihood ratio. The conditional KL chain rule therefore gives

$$\begin{aligned}\mathrm{KL}(P_d^\star \| P_d^{q^p}) &= \sum_{t=1}^{H} \mathbb{E}_{P_d^\star} \mathrm{KL}(q_t^\star(\cdot \mid H_t, A_t) \| q_t^p(\cdot \mid H_t, A_t)) \\ &\le \sum_{t=1}^{H} \mathbb{E}_{P_d^\star} \mathrm{KL}(p_t^\star(\cdot \mid H_t, A_t) \| p_t(\cdot \mid H_t, A_t))\,.\end{aligned}$$

Finally, suppose $p, p' \in \mathcal{P}_t$ satisfy

$$e^{-r} \le \frac{dp(\cdot \mid x)}{dp'(\cdot \mid x)} \le e^r \qquad p'(\cdot \mid x)\text{-a.s.}$$

Writing $T_x(\xi) = \Phi_t(x, \xi)$, for every measurable B,

$$q^p(B \mid x) = \int \mathbf{1}\{T_x(\xi) \in B\}\, p(d\xi \mid x) \le e^r q^{p'}(B \mid x),$$

and the reverse domination follows analogously. Hence $q^p \ll q^{p'} \ll q^p$ and

$$e^{-r} \le \frac{dq^p(\cdot \mid x)}{dq^{p'}(\cdot \mid x)} \le e^r \qquad q^{p'}(\cdot \mid x)\text{-a.s.}$$

Thus every r-cover of $\mathcal{P}_t$ pushes forward to an r-cover of $\mathcal{Q}_{\Phi,t}$, proving $N_{\log}(\mathcal{Q}_{\Phi,t}, r) \le N_{\log}(\mathcal{P}_t, r)$. □

C.3 Proof of Proposition 3

Proof. Under a common initial law and decision kernel, the initial and action factors cancel from the Radon–Nikodym derivative of the true trajectory law with respect to the learned trajectory law. The conditional KL chain rule therefore gives (17). Define on the time-indexed disjoint union

$$k(t, h, a) := \mathrm{KL}(q_t^\star(\cdot \mid h,a) \| q_t(\cdot \mid h,a)) \ge 0.$$

Then

$$\mathrm{KL}(P_d^\star \| P_d^q) = \int k\, d\rho_d^\star \le C_{\mathrm{cov}} \int k\, d\rho_\mu^\star = C_{\mathrm{cov}} \mathrm{KL}(P_\mu^\star \| P_\mu^q),$$

which proves (18). □

C.4 Proof of Theorem 2

Proof. A balanced Varshamov–Gilbert lemma gives universal constants $c_1, C_1 > 0$ such that, for every $M \geq 4$, there are an even integer $d \leq C_1 \log M$ and balanced vectors

$$v^{(1)}, \ldots, v^{(M)} \in \{-1, +1\}^d, \qquad \sum_{x=1}^{d} v_x^{(j)} = 0,$$

whose pairwise Hamming distances satisfy

$$D\big(v^{(j)}, v^{(k)}\big) \geq c_1 d, \qquad j \neq k.$$

This is the usual greedy packing of the middle slice of the binary cube; the middle slice has cardinality $\binom{d}{d/2}$ and a Hamming ball of any fixed relative radius occupies at most an exponentially smaller fraction of it.

Let $a > 0$ be a sufficiently small universal constant and set

$$\alpha = aG\sqrt{\frac{\log M}{n}}.$$

Increasing the universal constant c_0 in the theorem if necessary ensures $\alpha \leq \min\{1/4, G/4\}$. On the sample space $[d]$, define

$$q_j(x) := \frac{1 + \alpha v_x^{(j)}}{d}, \qquad j \in [M], \quad x \in [d], \tag{22}$$

and let $\mathcal{Q}_0 = \{q_1, \ldots, q_M\}$. Balance makes every q_j a probability density. For all j, k, x,

$$\left|\log \frac{q_j(x)}{q_k(x)}\right| \leq \log \frac{1+\alpha}{1-\alpha} \leq 3\alpha \leq G,$$

so the class satisfies Assumption 1 whichever member is the truth.

If $j \neq k$, balance implies that the two types of mismatched signs occur equally often. Therefore

$$\mathrm{KL}(q_j \| q_k) = \frac{D(v^{(j)}, v^{(k)})}{d}\, \alpha \log \frac{1+\alpha}{1-\alpha}. \tag{23}$$

Since $2\alpha \leq \log((1+\alpha)/(1-\alpha)) \leq 3\alpha$ for $\alpha \leq 1/4$, it follows that

$$2c_1\alpha^2 \leq \mathrm{KL}(q_j \| q_k) \leq 3\alpha^2, \qquad j \neq k. \tag{24}$$

We first consider one coordinate. Let J be uniform on $[M]$ and, given $J = j$, observe $Y^1, \ldots, Y^n \overset{\text{iid}}{\sim} q_j$. Convexity of KL in its second argument gives the standard mutual-information bound

$$I(J; Y^{1:n}) \leq \frac{1}{M^2} \sum_{j,k=1}^{M} \mathrm{KL}(q_j^{\otimes n} \| q_k^{\otimes n}) \leq 3n\alpha^2 \leq \frac{1}{4} \log M,$$

where the last inequality holds by choosing a sufficiently small and using $G \leq 1$. Fano's inequality then implies, for every possibly randomized estimator $\widehat{J}$,

$$\Pr(\widehat{J} \neq J) \geq 1 - \frac{I(J; Y^{1:n}) + \log 2}{\log M} \geq \frac{1}{4},$$

where the last step uses $M \geq 4$.

For the H-coordinate problem, draw $J_1, \ldots, J_H$ independently and uniformly from $[M]$, and put

$$P_J := \bigotimes_{h=1}^{H} q_{J_h}.$$

Write $\mathcal{D}_h = (Y_h^1, \ldots, Y_h^n)$ and $\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_H)$. Under the product prior and product experiment, the posterior factorizes:

$$\mathcal{L}(J \mid \mathcal{D}) = \bigotimes_{h=1}^{H} \mathcal{L}(J_h \mid \mathcal{D}_h).$$

We first lower-bound the unrestricted one-coordinate Bayes risk. Define the posterior mean codeword and posterior predictive density

$$m_h := \mathbb{E}[v^{(J_h)} \mid \mathcal{D}_h] \in [-1, 1]^d, \qquad \overline{q}_h(x) := \mathbb{E}[q_{J_h}(x) \mid \mathcal{D}_h] = \frac{1 + \alpha (m_h)_x}{d}.$$

Let $\widetilde{J}_h$ be a nearest Euclidean codeword to m_h, with measurable tie breaking. If $\widetilde{J}_h \neq J_h$, packing separation and nearestness give

$$2\sqrt{c_1 d} \leq \|v^{(\widetilde{J}_h)} - v^{(J_h)}\|_2 \leq 2\|m_h - v^{(J_h)}\|_2.$$

The preceding Fano bound applies to this decoder, so

$$\mathbb{E}\|v^{(J_h)} - m_h\|_2^2 \geq c_1 d \, \Pr(\widetilde{J}_h \neq J_h) \geq \frac{c_1 d}{4}. \tag{25}$$

For $f(p) = \sum_x p_x \log p_x$, every coordinate on the line segment from q_{J_h} to $\overline{q}_h$ is at most $(1 + \alpha)/d \leq 5/(4d)$; hence $\nabla^2 f \succeq (4d/5) I$ there. The Bregman-divergence identity for KL and strong convexity yield

$$\begin{aligned} \mathrm{KL}(q_{J_h} \| \overline{q}_h) &\geq \frac{2d}{5} \|q_{J_h} - \overline{q}_h\|_2^2 \\ &= \frac{2\alpha^2}{5d} \|v^{(J_h)} - m_h\|_2^2. \end{aligned} \tag{26}$$

Combining the last two displays gives

$$\mathbb{E}\mathrm{KL}(q_{J_h} \| \overline{q}_h) \geq \frac{c_1}{10} \alpha^2. \tag{27}$$

It remains to show that an arbitrary joint output cannot beat the sum of these posterior-predictive risks. Posterior factorization gives

$$\overline{P}_{\mathcal{D}} := \mathbb{E}[P_J \mid \mathcal{D}] = \bigotimes_{h=1}^{H} \overline{q}_h.$$

For every trajectory law Q, inserting $\overline{P}_{\mathcal{D}}$ in the log ratio and using $\mathbb{E}[P_J \mid \mathcal{D}] = \overline{P}_{\mathcal{D}}$ gives the conditional Bayes projection identity

$$\mathbb{E}[\mathrm{KL}(P_J \| Q) \mid \mathcal{D}] = \mathbb{E}[\mathrm{KL}(P_J \| \overline{P}_{\mathcal{D}}) \mid \mathcal{D}] + \mathrm{KL}(\overline{P}_{\mathcal{D}} \| Q). \tag{28}$$

The identity remains valid after averaging over estimator randomization. Therefore every unrestricted estimator $\widehat{P}(\mathcal{D})$ satisfies

$$\begin{aligned}\mathbb{E}\mathrm{KL}(P_J\|\widehat{P}(\mathcal{D})) &\geq \mathbb{E}\mathrm{KL}(P_J\|\overline{P}_{\mathcal{D}})\\ &= \sum_{h=1}^{H}\mathbb{E}\mathrm{KL}(q_{J_h}\|\overline{q}_h)\\ &\geq \frac{c_1}{10}H\alpha^2.\end{aligned}$$

The supremum risk is at least this Bayes risk. Substituting $\alpha = aG\sqrt{\log M/n}$ and absorbing $a^2c_1/10$ into c proves (10).

□

C.5 Proof of Theorem 3

Proof. Fix $M \geq 4$. A binary Gilbert–Varshamov packing gives a universal constant C_0 and an integer $d \leq C_0 \log M$ for which there are exactly M codewords

$$v^{(1)},\ldots,v^{(M)} \in \{-1,+1\}^d, \qquad D\big(v^{(k)},v^{(\ell)}\big) \geq \frac{d}{4} \quad (k \neq \ell),$$

where D is Hamming distance. Each trajectory begins with a common, model-independent observable context $X = (J, R)$, where

$$J \sim \mathrm{Unif}[d], \qquad R \sim \mathrm{Bernoulli}(\rho), \qquad \rho := \frac{\log M}{16n\log 2}.$$

The variables J and R are independent. Choosing the universal c_0 in the theorem sufficiently large ensures $\rho \leq 1$. There is a singleton action space, so no action factor enters any likelihood ratio.

Set

$$\alpha := \frac{1}{2}\tanh\left(\frac{G}{2}\right).$$

For each codeword $v^{(k)}$, define one history kernel q_k and use that *same kernel at every round*:

$$q_k(Y_h = 1 \mid X, Y_{<h}) = \frac{1}{2} + \alpha R v_J^{(k)}, \qquad h \in [H]. \tag{29}$$

Thus $Y_1,\ldots,Y_H$ are conditionally independent given X, and the kernel ignores previous outcomes. This is a genuinely fully shared class of exactly M elements, not a product class with independently selected stagewise indices. The context is an initial observable with a common marginal and hence cancels from every KL ratio.

Let $p_\pm = 1/2 \pm \alpha$. By the definition of α,

$$\log\frac{p_+}{p_-} = G.$$

Consequently, all conditional log ratios between members of the class have absolute value at most G. The KL divergence between the opposite-sign Bernoulli laws, in either direction, is

$$\kappa_G := \mathrm{KL}(\mathrm{Bern}(p_+)\,\|\,\mathrm{Bern}(p_-)) = 2\alpha G = G\tanh\left(\frac{G}{2}\right) \geq \frac{G^2}{4}, \tag{30}$$

where the final inequality uses $G \le 1$. Averaging first over the H continuations and then over X gives, for $k \neq \ell$,

$$\begin{aligned} \mathrm{KL}(P_k \| P_\ell) &= \frac{\rho H \kappa_G}{d} D\big(v^{(k)}, v^{(\ell)}\big) \\ &\ge \frac{\rho H \kappa_G}{4}. \end{aligned} \tag{31}$$

It remains to show that the training sample cannot reliably identify the codeword. Put the uniform prior on $K \in [M]$ and let $\mathcal{D}_n = (X_i, Y_{i,1:H})_{i=1}^n$. The contexts are independent of K. If $R_i = 0$, the corresponding continuation has no information about K; if $R_i = 1$, its law depends on K only through the single binary code bit $v_{J_i}^{(K)}$. The data-processing inequality and the chain rule for mutual information therefore give

$$I(K; \mathcal{D}_n) \le \sum_{i=1}^{n} \rho \log 2 = n\rho \log 2 = \frac{1}{16} \log M.$$

This argument remains valid even when H is arbitrarily large: the H repeated observations can reveal the activated code bit accurately, but one activated context carries at most one bit of information about the codeword. Fano's inequality now implies, for every possibly randomized estimator $\widehat{K}$,

$$\Pr(\widehat{K} \neq K) \ge 1 - \frac{I(K; \mathcal{D}_n) + \log 2}{\log M} \ge \frac{7}{16}.$$

Every proper $\mathcal{Q}_{\mathrm{sh}}$-valued estimator has the form $\hat{q} = q_{\widehat{K}}$. Combining the last display with (31) yields the Bayes-risk bound

$$\mathbb{E}\mathrm{KL}(P_K \| P_{\widehat{K}}) \ge \frac{7}{64} \rho H \kappa_G \ge c \, \frac{H G^2 \log M}{n}$$

for a universal $c > 0$. The supremum risk is at least the Bayes risk, proving (11). The least-favorable activation probability $\rho \asymp \log M / n$ depends on (M, n), as is standard for a local minimax construction. □

Why the original product proof cannot simply be tied. On the diagonal of the original construction, the same index is seen at all H rounds. If its per-round separation is ϵ^2, then the test trajectory loss is of order $H\epsilon^2$, but the information in n training trajectories is of order $nH\epsilon^2$. Fano therefore requires $\epsilon^2 \asymp \log M/(nH)$ and yields only $H\epsilon^2 \asymp \log M/n$. The rare-context construction above is different: an activated training trajectory conveys at most one code bit, whereas choosing the wrong proper shared kernel incurs that error at all H test rounds.

Why the proper-learning qualifier is essential. Proposition 2 sharpens the static uniform-mixture observation by incorporating the sample size: progressive posterior aggregation over these M shared trajectory laws has realizable expected KL at most $\log M/(n+1)$, independently of H. Thus the shared lower bound cannot be a theorem for arbitrary improper trajectory mixtures; it is a proper-selection lower bound for ERM and other selectors required to return one member of $\mathcal{Q}_{\mathrm{sh}}$.

C.6 Proof of Proposition 2

Proof. Let $p_1, \dots, p_N$ be densities with respect to a common dominating measure. Given training trajectories $Z_1, \dots, Z_n$, define, for $t = 0, \dots, n$,

$$w_{k,t} := \frac{\prod_{i=1}^{t} p_k(Z_i)}{\sum_{\ell=1}^{N} \prod_{i=1}^{t} p_\ell(Z_i)}, \qquad Q_t := \sum_{k=1}^{N} w_{k,t} P_k,$$

with $w_{k,0} = 1/N$, and set

$$\widehat{P}_{\mathrm{mix}} := \frac{1}{n+1} \sum_{t=0}^{n} Q_t.$$

Thus Q_t is the uniform-prior posterior predictive law after the first t training trajectories, and the displayed average is a deterministic estimator based on the full sample. On a zero-denominator event the weights may be defined arbitrarily; whenever the oracle term is finite, such events are $P^\star$-null.

Write q_t for the density of Q_t. The predictive likelihoods telescope:

$$\prod_{t=0}^{n} q_t(Z_{t+1}) = \frac{1}{N} \sum_{\ell=1}^{N} \prod_{i=1}^{n+1} p_\ell(Z_i).$$

Consequently, for every fixed $k \in [N]$ and every realization of $Z_{1:n+1}$,

$$\sum_{t=0}^{n} \log \frac{p_k(Z_{t+1})}{q_t(Z_{t+1})} \le \log N, \tag{32}$$

because the mixture on the right contains the kth likelihood with weight $1/N$.

Now let $Z_1, \dots, Z_{n+1}$ be iid from an arbitrary $P^\star$. Conditional expectation given $Z_{1:t}$ turns the tth expected log ratio in (32) into

$$\mathbb{E}\mathrm{KL}(P^\star \| Q_t) - \mathrm{KL}(P^\star \| P_k).$$

Taking expectations, summing, and using convexity of KL in its second argument therefore gives

$$\begin{aligned} \mathbb{E}\mathrm{KL}(P^\star \| \widehat{P}_{\mathrm{mix}}) &\le \frac{1}{n+1} \sum_{t=0}^{n} \mathbb{E}\mathrm{KL}(P^\star \| Q_t) \\ &\le \mathrm{KL}(P^\star \| P_k) + \frac{\log N}{n+1}. \end{aligned}$$

Minimizing over k proves (12). For a realizable decomposable family $N = M^H$, the remainder is $H \log M/(n+1)$; for a realizable fully shared family $N = M$, it is $\log M/(n+1)$. □

C.7 Proof of Theorem 4

Proof. We first record the likelihood-ratio variance inequality used below. Let X be a random variable such that $|X| \le B$ and $\mathbb{E}e^{-X} \le 1$. Then

$$\mathbb{E}X^2 \le 3(B+1)\mathbb{E}X. \tag{33}$$

Indeed, set $\psi(x) = e^{-x} - 1 + x$. For $x \in [-B, 0]$, $\psi(x) \geq x^2/2$; for $x \in [0, 1]$, $\psi(x) \geq x^2/3$; and for $x \in [1, B]$, $\psi(x) \geq x/e$. Hence $x^2 \leq 3(B+1)\psi(x)$ throughout $[-B, B]$. Moreover,

$$\mathbb{E}\psi(X) = \mathbb{E}e^{-X} - 1 + \mathbb{E}X \leq \mathbb{E}X,$$

which proves (33). In particular, $\mathbb{E}X \geq 0$ and $\mathrm{Var}(X) \leq 3(B+1)\mathbb{E}X$.

We next derive a finite-class relative-deviation event. Let $X_1, \ldots, X_n$ be i.i.d. copies of such an X, write $K = \mathbb{E}X$, and let $P_n X = n^{-1}\sum_i X_i$. Two-sided Bernstein's inequality and (33) imply that, with probability at least $1 - 2e^{-x}$,

$$\begin{aligned} |(P_n - P)X| &\leq \sqrt{\frac{6(B+1)Kx}{n}} + \frac{2Bx}{3n} \\ &\leq \eta K + C_0 \frac{(B+1)x}{\eta n}, \qquad 0 < \eta < 1, \end{aligned} \tag{34}$$

for a universal constant C_0. The final line follows from Young's inequality.

For the decomposable class, define

$$X_{h,q}(U_{1:H}) := \log \frac{q_h^\star(U_h \mid U_{<h})}{q(U_h \mid U_{<h})}, \qquad K_h(q) := \mathbb{E}_{P^\star} X_{h,q}.$$

The conditional-density normalization gives

$$\mathbb{E}_{P^\star} e^{-X_{h,q}} = \mathbb{E}_{P^\star}\left[\int_{\mathrm{supp}(q_h^\star(\cdot|U_{<h}))} q(du \mid U_{<h})\right] \leq 1.$$

Assumption 1 gives $|X_{h,q}| \leq G$. Apply (34) to every $(h, q) \in [H] \times \mathcal{Q}_0$ with

$$x = \log \frac{2H|\mathcal{Q}_0|}{\delta}.$$

A union bound shows that, with probability at least $1 - \delta$, simultaneously for all h and q,

$$|(P_n - P)X_{h,q}| \leq \eta K_h(q) + R, \qquad R := C_0 \frac{(G+1)x}{\eta n}. \tag{35}$$

Because the empirical objective separates across coordinates, each component $\hat{q}_h$ is an empirical minimizer over $\mathcal{Q}_0$. Let $q_h^\circ \in \mathrm{argmin}_{q \in \mathcal{Q}_0} K_h(q)$, which exists because the class is finite. The $q_h^\star$ term is common to every empirical loss, so $P_n X_{h,\hat{q}_h} \leq P_n X_{h,q_h^\circ}$. On the event in (35),

$$(1 - \eta)K_h(\hat{q}_h) \leq (1 + \eta)K_h(q_h^\circ) + 2R.$$

Take $\eta = \varepsilon/(2 + \varepsilon)$, for which $(1 + \eta)/(1 - \eta) = 1 + \varepsilon$. Summing over h, using the exact KL chain rule, and absorbing numerical factors into a universal C gives (14).

For the fully shared class, define on an entire trajectory

$$X_q(U_{1:H}) := \sum_{h=1}^{H} \log \frac{q_h^\star(U_h \mid U_{<h})}{q(U_h \mid U_{<h})} = \log \frac{dP^\star}{dP_q}(U_{1:H}).$$

Here $|X_q| \leq HG$ and $\mathbb{E}_{P^\star} e^{-X_q} \leq 1$; equality holds when P_q assigns no mass outside the support of $P^\star$. Apply (34) with $B = HG$ to all $q \in \mathcal{Q}_0$ and $x = \log(2|\mathcal{Q}_0|/\delta)$. The same empirical-optimality argument, now for the single shared component, yields

$$(1 - \eta)\mathrm{KL}(P^\star \| P_{\hat{q}}) \leq (1 + \eta) \inf_{q \in \mathcal{Q}_{\mathrm{sh}}} \mathrm{KL}(P^\star \| P_q) + C_0 \frac{(HG+1)x}{\eta n}.$$

Taking $\eta = \varepsilon/(2 + \varepsilon)$ proves (15). □

C.8 Proof of Theorem 5

Proof. Let $q^\circ = (q_1^\circ, \dots, q_H^\circ) \in \mathcal{Q}_0^H$. For each h, choose $\tilde{q}_h \in \mathcal{Q}_{0,r}$ such that

$$d_{\log}(q_h^\circ, \tilde{q}_h) \le r.$$

Then

$$|K_h(\tilde{q}_h) - K_h(q_h^\circ)| = \left| \mathbb{E}_{P^\star} \log \frac{q_h^\circ(U_h \mid U_{<h})}{\tilde{q}_h(U_h \mid U_{<h})} \right| \le r.$$

Summing and using the KL chain rule yields

$$\inf_{\tilde{q} \in \mathcal{Q}_{0,r}^H} \mathrm{KL}(P^\star \| P_{\tilde{q}}) \le \inf_{q \in \mathcal{Q}_0^H} \mathrm{KL}(P^\star \| P_q) + Hr.$$

Moreover, the triangle inequality for log ratios shows that every net element has true-to-model log ratio bounded by $G + r$. Applying Theorem 4 to the finite class $\mathcal{Q}_{0,r}$ and absorbing the resulting $(1+\varepsilon)Hr$ term into $C'Hr$ proves the decomposable claim. The shared-class argument is identical, using one common net element across all horizons. □

C.9 Proof of Theorem 6

Proof. Part 1: generic bounded objectives. By optimality of $\hat{d}$ under the learned law, $J_{\hat{q}}(d^\star) - J_{\hat{q}}(\hat{d}) \le 0$. Therefore

$$\begin{aligned} J^\star(d^\star) - J^\star(\hat{d}) &= \{J^\star(d^\star) - J_{\hat{q}}(d^\star)\} + \{J_{\hat{q}}(d^\star) - J_{\hat{q}}(\hat{d})\} + \{J_{\hat{q}}(\hat{d}) - J^\star(\hat{d})\} \\ &\le |J^\star(d^\star) - J_{\hat{q}}(d^\star)| + |J_{\hat{q}}(\hat{d}) - J^\star(\hat{d})| \\ &\le 2 \sup_{d \in D(h)} |J^\star(d) - J_{\hat{q}}(d)|. \end{aligned}$$

We use the convention $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$. For fixed d, subtracting a constant from F_d does not change the difference of its expectations. The variational characterization of total variation therefore gives

$$\left| \mathbb{E}_{P_d^\star} F_d - \mathbb{E}_{P_d^{\hat{q}}} F_d \right| \le \mathrm{range}(F_d)\, \mathrm{TV}(P_d^\star, P_d^{\hat{q}}) \le B \mathrm{TV}(P_d^\star, P_d^{\hat{q}}).$$

Taking the supremum over d gives the corresponding total-variation bound. Finally, Pinsker's inequality gives

$$\mathrm{TV}(P_d^\star, P_d^{\hat{q}}) \le \sqrt{\frac{1}{2} \mathrm{KL}(P_d^\star \| P_d^{\hat{q}})},$$

which proves (19).

Part 2: worst-case sharpness. For part 2, fix sufficiently small $\kappa > 0$ and set

$$\varepsilon := \frac{1}{2} \sqrt{1 - e^{-2\kappa}}.$$

Consider two decisions, S and R. Under S, the true and learned continuation laws coincide and the objective is the constant

$$F_S \equiv B \left(\frac{1}{2} + \frac{\varepsilon}{2} \right).$$

Under R, let $X \in \{0,1\}$, set $F_R(X) = BX$, and take

$$X \sim \mathrm{Bern}\left(\frac{1}{2}\right) \quad \text{under the true law,} \qquad X \sim \mathrm{Bern}\left(\frac{1}{2}+\varepsilon\right) \quad \text{under the learned law.}$$

Both objectives have range at most B, and

$$\Delta_D(h;\hat{q}) = \mathrm{kl}\left(\frac{1}{2}\,\middle\|\,\frac{1}{2}+\varepsilon\right) = -\frac{1}{2}\log(1-4\varepsilon^2) = \kappa.$$

The true problem uniquely prefers S, whereas the learned problem uniquely prefers R. The true plug-in regret is

$$B\left(\frac{1}{2}+\frac{\varepsilon}{2}\right) - \frac{B}{2} = \frac{B\varepsilon}{2}.$$

For $0 < \kappa \le 1/2$, the elementary bound $1 - e^{-2\kappa} \ge \kappa$ gives

$$\frac{B\varepsilon}{2} = \frac{B}{4}\sqrt{1-e^{-2\kappa}} \ge \frac{B}{4}\sqrt{\kappa}.$$

Thus part 2 holds with $c = 1/4$.

Part 3: fast transfer. Strong concavity of $J^\star$, applied with $x = \hat{d}$ and $y = d^\star$,

$$J^\star(d^\star) - J^\star(\hat{d}) \le \langle \nabla J^\star(\hat{d}), d^\star - \hat{d}\rangle - \frac{\mu}{2}\|d^\star - \hat{d}\|^2.$$

Because $D(h)$ is convex and $\hat{d}$ globally maximizes the differentiable function $J_{\hat{q}}$, the one-dimensional function

$$t \longmapsto J_{\hat{q}}\big(\hat{d} + t(d^\star - \hat{d})\big), \qquad t \in [0,1],$$

is maximized at $t = 0$. Its right derivative is therefore nonpositive:

$$\langle \nabla J_{\hat{q}}(\hat{d}), d^\star - \hat{d}\rangle \le 0.$$

Consequently, writing

$$\eta := \sup_{d \in D(h)} \|\nabla J_{\hat{q}}(d) - \nabla J^\star(d)\|_*,$$

duality of the norms yields

$$\begin{aligned} J^\star(d^\star) - J^\star(\hat{d}) &\le \left\langle \nabla J^\star(\hat{d}) - \nabla J_{\hat{q}}(\hat{d}), d^\star - \hat{d}\right\rangle - \frac{\mu}{2}\|d^\star - \hat{d}\|^2 \\ &\le \eta\|d^\star - \hat{d}\| - \frac{\mu}{2}\|d^\star - \hat{d}\|^2 \le \frac{\eta^2}{2\mu}. \end{aligned}$$

The gradient-stability assumption gives $\eta^2 \le L_{\mathrm{KL}}^2 \Delta_D(h;\hat{q})$, proving the claim.

Part 4: objective reuse. For part 4, fix any family $\{F_d : d \in D(h)\}$ satisfying the stated uniform range bound. The induced laws $P_d^\star$ and $P_d^{\hat{q}}$ do not depend on which downstream objective is applied. Part 1 therefore yields

$$J_F^\star(d_F^\star) - J_F^\star(\hat{d}_F) \le B_F\sqrt{2\Delta_D(h;\hat{q})} \le B_F\sqrt{2\kappa}.$$

This proves the objective-reuse claim. □

C.10 Proof of Theorem 7

Proof. Because $\mathcal{Q} = \{\bar{q}\}$, the quantity calculated below is

$$\inf_{q \in \mathcal{Q}} \sup_{\pi} \mathrm{KL}(P^{\pi}_{q^\star} \| P^{\pi}_{q});$$

it is a class-approximation term, not an estimation error.

The construction is quite intuitive. Consider a finite-horizon problem with one initial state s_0, absorbing states C, G, B, and two actions: a safe action S and a risky action R. After the first action, the system enters one of three absorbing states C, G, B, and remains there for H reward periods. State C yields reward

$$c := \frac{1}{2} + \frac{\varepsilon}{2}$$

per period, state G yields reward 1 per period, and state B yields reward 0 per period.

The safe action S deterministically sends the system to C under both $q^\star$ and $\bar{q}$. The risky action R sends the system to B with probability $1/2$ under the true law $q^\star$, and with probability $1/2 - \varepsilon$ under the misspecified law $\bar{q}$. Thus

$$q^\star(B \mid s_0, R) = \frac{1}{2}, \qquad \bar{q}(B \mid s_0, R) = \frac{1}{2} - \varepsilon.$$

All other transitions are deterministic and identical under $q^\star$ and $\bar{q}$.

We first calculate the continuation-law misspecification. If a policy chooses R with probability α at s_0, the common decision kernel cancels and the only model discrepancy is the transition from s_0. Hence

$$\mathrm{KL}(P^{\pi}_{q^\star} \| P^{\pi}_{\bar{q}}) = \alpha \kappa_\varepsilon,$$

where

$$\begin{aligned} \kappa_\varepsilon &:= \mathrm{kl}\left(\frac{1}{2} \,\middle\|\, \frac{1}{2} - \varepsilon\right) \\ &= -\frac{1}{2}\log(1 - 4\varepsilon^2). \end{aligned}$$

The supremum is attained by the deterministic risky policy. Since $x \le -\log(1-x) \le x/(1-x)$ for $x \in [0, 1)$, taking $x = 4\varepsilon^2 \le 1/4$ gives

$$2\varepsilon^2 \le \kappa_\varepsilon \le \frac{8}{3}\varepsilon^2.$$

Now compare the optimal decisions. Under the true law,

$$Q^\star_{q^\star,0}(s_0, S) = H\left(\frac{1}{2} + \frac{\varepsilon}{2}\right),$$

while

$$Q^\star_{q^\star,0}(s_0, R) = H \cdot \frac{1}{2}.$$

Hence the true optimal action is S.

Under the misspecified law $\bar{q}$,

$$Q^\star_{\bar{q},0}(s_0, S) = H\left(\frac{1}{2} + \frac{\varepsilon}{2}\right),$$

but

$$Q^\star_{\bar{q},0}(s_0, R) = H\left(\frac{1}{2} + \varepsilon\right).$$

Therefore the model-optimal action is R. The policy transferred from the misspecified continuation law is thus wrong under the true model, and its true regret is

$$Q^\star_{q^\star,0}(s_0, S) - Q^\star_{q^\star,0}(s_0, R) = \frac{H\varepsilon}{2}.$$

For the value approximation, the true optimal value at s_0 is

$$V^\star_{q^\star,0}(s_0) = H\left(\frac{1}{2} + \frac{\varepsilon}{2}\right),$$

whereas the misspecified optimal value is

$$V^\star_{\bar{q},0}(s_0) = H\left(\frac{1}{2} + \varepsilon\right).$$

Thus

$$\|V^\star_{\bar{q}} - V^\star_{q^\star}\|_\infty \geq \frac{H\varepsilon}{2}.$$

Similarly,

$$\left|Q^\star_{\bar{q},0}(s_0, R) - Q^\star_{q^\star,0}(s_0, R)\right| = H\varepsilon,$$

so

$$\|Q^\star_{\bar{q}} - Q^\star_{q^\star}\|_\infty \geq H\varepsilon.$$

Finally, $\kappa_\varepsilon \leq (8/3)\varepsilon^2$ implies

$$\frac{H\varepsilon}{2} \geq H\sqrt{\frac{3}{32}}\sqrt{\kappa_\varepsilon}.$$

Thus the transferred policy and value errors are $\Omega(H\sqrt{\kappa_\varepsilon})$, while $\kappa_\varepsilon = \Theta(\varepsilon^2)$. □

Mechanism, normalization, and scope. The likelihood ratio is incurred once, when the absorbing state is selected; the subsequent deterministic persistence reveals no new model discrepancy. This is why the full observable-trajectory KL remains $\Theta(\varepsilon^2)$ independently of H. The one selected state determines all H payoffs, so the induced policy and value errors scale with the objective range. Dividing rewards by H removes this scale factor but leaves normalized policy and value error at least $\sqrt{3/32}\sqrt{\kappa_\varepsilon}$, compared with trajectory error κ_ε.

For completeness, the strongly curved companion uses a persistent $Y \sim \mathrm{Bern}(1/2)$ under truth and $Y \sim \mathrm{Bern}(1/2 + \delta)$ under the model, with total expected profit

$$J^\star_H(p) = Hp(9/2 - p), \qquad \bar{J}_H(p) = Hp(9/2 + \delta - p), \qquad p \in [1, 3].$$

The observation-history KL is $\kappa_\delta = -\frac{1}{2}\log(1 - 4\delta^2) = \Theta(\delta^2)$, the negative objectives are $2H$-strongly convex, and the misspecified optimizer moves by $\delta/2$. Its exact true regret is therefore $H\delta^2/4 = \Theta(H\kappa_\delta)$. Restricting the same ambient quadratic to the prespecified two-price near-tie makes the argmax switch discontinuously and gives $H\delta/2 = \Theta(H\sqrt{\kappa_\delta})$ regret. Strong curvature plus gradient stability thus controls decision regret; it does not make arbitrary value-level errors linear in KL. These examples diagnose the metric and geometry that a realizability-robust guarantee must specify, rather than proving a universal impossibility theorem.

The curvature in this calculation is entirely curvature of expected profit as a function of the chosen price. It neither assumes nor implies that an autoregressive neural likelihood is strongly convex in its weights. For a general autoregressive parameterization, one may verify the KL-to-gradient condition directly in prediction space or derive it locally from identifiable Fisher curvature, as explained in Remark 1.

D Experimental Design and Reproducibility

This appendix describes the experimental study at the level needed for exact replication. The experiments are mechanism studies rather than a broad forecasting benchmark. The finite-class experiment isolates the local H/n joint-KL mechanism; the inventory and queueing experiments examine structural feasibility, censored likelihoods, KL contraction, and plug-in decisions; the pricing experiment separates worst-case horizon-sensitive transfer from fast transfer under strong concavity; and the Rossmann study introduces real serial and calendar variation into a clearly labeled semi-synthetic inventory system.

D.1 Generic structural likelihood

In the structural experiments, the latent shock ξ_{t+1} has finite support $\Xi = \{\xi_1, \dots, \xi_K\}$, and the next observable continuation is generated by a deterministic reactor

$$Z_{t+1} = \Phi(H_t, A_t, \xi_{t+1}).$$

A neural sequence model produces a categorical shock law

$$p_\theta(\xi_k \mid H_t, A_t) = \frac{\exp\{g_{\theta,k}(H_t, A_t)\}}{\sum_{\ell=1}^{K} \exp\{g_{\theta,\ell}(H_t, A_t)\}}.$$

For an observed continuation z_{t+1}, define its latent preimage

$$\mathcal{E}_t(z_{t+1}) := \{k : \Phi(H_t, A_t, \xi_k) = z_{t+1}\}.$$

The structural models are trained using the exact observed-data negative log-likelihood

$$\ell_t(\theta) = -\log \sum_{k \in \mathcal{E}_t(Z_{t+1})} p_\theta(\xi_k \mid H_t, A_t). \tag{36}$$

We evaluate the sum in (36) using a masked `logsumexp`. Thus a censored demand or reflected net input is treated as an event, not replaced by an artificial point label.

The induced observable law is the finite pushforward

$$q_\theta(z \mid h, a) = \sum_{k:\Phi(h,a,\xi_k)=z} p_\theta(\xi_k \mid h, a).$$

Consequently, every structural draw lies in

$$\mathcal{S}(h, a) := \{\Phi(h, a, \xi) : \xi \in \Xi\}.$$

The direct GRU baseline instead predicts a global categorical law over observable continuations without receiving the context-dependent support $\mathcal{S}(h, a)$. We also report a prespecified projection baseline,

$$q_{\mathrm{proj}}(z \mid h, a) = \frac{q_{\mathrm{dir}}(z \mid h, a)\mathbf{1}\{z \in \mathcal{S}(h, a)\}}{\sum_{z' \in \mathcal{S}(h,a)} q_{\mathrm{dir}}(z' \mid h, a)}. \tag{37}$$

Projection repairs feasibility but does not recover a reusable law for the latent operational shock.

D.2 Finite-class joint-KL mechanism

For each sample size n, define

$$\delta_n = \frac{0.5}{\sqrt{n}}, \qquad \mathcal{P}_n = \left\{ \frac{1}{2} - 3\delta_n,\ \frac{1}{2} - \delta_n,\ \frac{1}{2} + \delta_n,\ \frac{1}{2} + 3\delta_n \right\}.$$

At each coordinate $h \in [H]$, a true Bernoulli parameter $p_h^\star$ is drawn uniformly from $\mathcal{P}_n$, independently across coordinates. We then generate

$$N_h \sim \text{Binomial}(n, p_h^\star)$$

and choose the coordinate-wise maximum-likelihood estimate

$$\widehat{p}_h \in \operatorname*{argmin}_{p \in \mathcal{P}_n} \{ -N_h \log p - (n - N_h) \log(1 - p) \}.$$

The reported trajectory risk is

$$\sum_{h=1}^{H} \text{KL}(\text{Bern}(p_h^\star) \,\|\, \text{Bern}(\widehat{p}_h)).$$

The grid is

$$H \in \{5, 10, 20, 40, 80\}, \qquad n \in \{100, 250, 500, 1000, 2000\},$$

with 1,000 Monte Carlo replications per cell and master seed 20260714. We record the mean joint KL, its Monte Carlo standard error, the coordinate-selection error, the maximum log ratio in the local class, and the normalized quantity $n\,\text{KL}/H$. Because $\mathcal{P}_n$ is a local family whose spacing changes with n, this is a diagnostic of the lower-bound mechanism, not a claim that every shared neural model must exhibit the same learning curve.

D.3 Lost-sales inventory experiment

Data-generating process. Each simulated trajectory has 30 periods. Inventory and order quantities are integer-valued, with

$$I_t \in \{0, \ldots, 10\}, \qquad A_t \in \{0, \ldots, \min(6, 10 - I_t)\}.$$

The order arrives immediately, so available inventory is $Y_t = I_t + A_t$. Latent demand has support $\mathcal{D} = \{0, \ldots, 10\}$ and follows

$$D_{t+1} \mid H_t \sim \text{Binomial}(10, \pi_t).$$

Let $\sigma(u) = (1+e^{-u})^{-1}$, and let $[u]_a^b = \min\{\max\{u, a\}, b\}$. An observable covariate evolves according to

$$X_t = [\,0.72 X_{t-1} + 0.70 \varepsilon_t\,]_{-2.5}^{2.5}, \qquad \varepsilon_t \overset{\text{iid}}{\sim} N(0, 1),$$

with a Gaussian initialization. Promotion satisfies

$$P_t \sim \text{Bernoulli}\left(0.22 + 0.10\, \mathbf{1}\{t \bmod 7 \in \{4, 5\}\}\right).$$

Writing

$$s_t = \sin(2\pi t/7), \qquad c_t = \cos(2\pi t/7),$$

the demand parameter is

$$\eta_t = -0.95 + 0.46 s_t + 0.20 c_t + 0.32 X_t + 0.50 P_t + 0.95 M_t, \tag{38}$$

$$\pi_t = [\,\sigma(\eta_t)\,]_{0.05}^{0.95}. \tag{39}$$

Here M_t is an observable-history summary initialized at $M_0 = 0.30$ and updated only from censored sales:

$$M_{t+1} = 0.85 M_t + 0.15\,\frac{S_{t+1}}{10}.$$

The learner is not supplied with latent demand when a stockout occurs.

The lost-sales reactor is

$$S_{t+1} = \min\{D_{t+1}, Y_t\}, \qquad I_{t+1} = Y_t - S_{t+1}. \tag{40}$$

Initial inventory is uniform on $\{2, \ldots, 8\}$. The boundary indicator used as a lagged feature is

$$B_{t+1}^I = \mathbf{1}\{S_{t+1} = Y_t\}.$$

This indicator records that the observation lies at the censoring boundary; in particular, when $Y_t = 0$, it should not be interpreted as proof of strictly positive lost demand.

Logging policy. The observable-history base-stock target is

$$T_t = \left[\text{round}\,(3 + 2.2 M_t + 1.4 P_t + 0.35 X_t)\right]_0^{10}.$$

Let

$$u_t = \min\{6, 10 - I_t\}, \qquad b_t = [\,T_t - I_t\,]_0^{u_t},$$

and define the unique clipped neighborhood

$$\mathcal{C}_t = \text{unique}\,\{[\,b_t - 1\,]_0^{u_t}, b_t, [\,b_t + 1\,]_0^{u_t}\}.$$

The logging law is

$$\mu_t(a \mid H_t) = \frac{0.35}{u_t + 1} + 0.65\,\frac{\mathbf{1}\{a \in \mathcal{C}_t\}}{|\mathcal{C}_t|}, \qquad a \in \{0, \ldots, u_t\}. \tag{41}$$

Thus every feasible order has positive probability, and the logging policy uses neither current nor future demand.

Observed likelihood and features. For an observation $(S_{t+1}, I_{t+1}) = (s, i')$, the demand preimage is

$$\mathcal{E}_t(s, i') = \begin{cases} \{s\}, & s < Y_t, \\ \{d \in \mathcal{D} : d \geq Y_t\}, & s = Y_t. \end{cases}$$

The structural likelihood is therefore the exact demand probability mass when $s < Y_t$ and the exact survival probability $\Pr_\theta(D_{t+1} \geq Y_t \mid H_t, A_t)$ at the censoring boundary.

The current feature token is

$$U_t^I = (I_t, A_t, Y_t, S_t, B_t^I, X_t, P_t, \\ \sin(2\pi t/7), \cos(2\pi t/7), M_t).$$

The direct GRU predicts one of $11^2 = 121$ ordered pairs (S_{t+1}, I_{t+1}), encoded as $11 S_{t+1} + I_{t+1}$. Its feasible support at a context with available inventory Y_t is

$$\mathcal{S}_I(H_t, A_t) = \{(s, Y_t - s) : s = 0, \ldots, Y_t\}.$$

Plug-in ordering criterion. For each held-out context, the candidate-action set is

$$\mathcal{A}_t^I = \{0, \dots, \min(6, 10 - I_t)\}.$$

The one-period observable cost is

$$C_I(I_{t+1}) = 0.35 I_{t+1} + 2\,\mathbf{1}\{I_{t+1} = 0\}. \tag{42}$$

For a candidate action, only the current action and available-inventory entries of the final causal token are changed. The structural model evaluates (42) under the pushforward of its learned demand law. The raw direct model evaluates the same cost under its global pair distribution; the projection baseline first applies (37). The oracle uses the known conditional demand law in (39). Reported decision regret is the oracle expected cost of the plug-in action minus the minimum oracle expected cost.

D.4 Controlled Lindley queue

Data-generating process. The queue length satisfies $Q_t \in \mathbb{Z}_+$, staffing actions belong to $\mathcal{A} = \{0, 1, 2, 3, 4\}$, and the latent net-input shock has support

$$\mathcal{X} = \{-4, -3, \dots, 4\}.$$

The observable covariate follows

$$X_t^e = [\,0.68 X_{t-1}^e + 0.75 \varepsilon_t\,]_{-2.5}^{2.5}, \qquad \varepsilon_t \overset{\text{iid}}{\sim} N(0,1).$$

Conditional on history and action, the net-input distribution is

$$p_t^\star(x \mid H_t, A_t) = \frac{\exp\left\{-\frac{1}{2}\left(\frac{x - L_t}{1.30}\right)^2\right\}}{\displaystyle\sum_{v=-4}^{4} \exp\left\{-\frac{1}{2}\left(\frac{v - L_t}{1.30}\right)^2\right\}}, \tag{43}$$

where

$$L_t = 0.80 + 0.55 \sin(2\pi t/12) + 0.28 X_t^e + 0.45 M_t^Q - 0.85 A_t.$$

The queue EWMA is initialized at $M_0^Q = 0.15$ and updated by

$$M_{t+1}^Q = 0.85 M_t^Q + 0.15 \min\{Q_{t+1}/12, 2\}.$$

Initial queue length is uniform on $\{0, \dots, 4\}$.

The controlled Lindley reactor is

$$Q_{t+1} = [\,Q_t + X_{t+1}\,]^+. \tag{44}$$

We record the observable increment

$$\Delta Q_{t+1} = Q_{t+1} - Q_t.$$

No exact-hit/overshoot indicator is supplied to the learner: when $Q_{t+1} = 0$, that distinction depends on the latent X_{t+1} and is not measurable from the observed queue transition.

Logging policy. Let

$$b_t = \big[\text{round}\{1 + 0.45Q_t + 0.60[X_t^e]^+\}\big]_0^4$$

and

$$\mathcal{C}_t^Q = \text{unique}\{[b_t - 1]_0^4,\, b_t,\, [b_t + 1]_0^4\}.$$

The action law is

$$\mu_t(a \mid H_t) = \frac{0.35}{5} + 0.65\,\frac{\mathbf{1}\{a \in \mathcal{C}_t^Q\}}{|\mathcal{C}_t^Q|}, \qquad a \in \{0, \ldots, 4\}. \tag{45}$$

Every staffing action therefore has positive logging probability.

Observed likelihood and features. When $Q_{t+1} > 0$, the shock is observed exactly: $X_{t+1} = Q_{t+1} - Q_t$. At the reflecting boundary,

$$Q_{t+1} = 0 \quad \Longleftrightarrow \quad X_{t+1} \le -Q_t.$$

Hence

$$\mathcal{E}_t(Q_{t+1}) = \begin{cases} \{Q_{t+1} - Q_t\}, & Q_{t+1} > 0, \\ \{x \in \mathcal{X} : x \le -Q_t\}, & Q_{t+1} = 0. \end{cases}$$

The structural queue model is trained using the corresponding exact event probability.

The feature token is

$$U_t^Q = \big(Q_t, A_t, \Delta Q_t, X_t^e, \sin(2\pi t/12), \cos(2\pi t/12), M_t^Q\big).$$

The direct GRU predicts Q_{t+1} on a global categorical grid $\{0, \ldots, K_Q - 1\}$, where the implementation sets

$$K_Q = \max_{\text{training transitions}} Q_{t+1} + 6.$$

This produced $K_Q \in \{16, \ldots, 19\}$ across the five archived synthetic seeds. At a given context its feasible support is

$$\mathcal{S}_Q(H_t, A_t) = \{[Q_t + x]^+ : x = -4, \ldots, 4\}.$$

Plug-in staffing criterion. The one-step staffing objective is

$$C_Q(a) = 0.30a + \mathbb{E}[Q_{t+1} \mid H_t, a]. \tag{46}$$

For counterfactual action a, the true shock location is recomputed as

$$L_t(a) = L_t(A_t) + 0.85A_t - 0.85a,$$

holding the observable history fixed. The learned models are evaluated by replacing the current staffing entry of the final token. Regret is the true objective of the plug-in staffing action minus $\min_{a \in \{0,\ldots,4\}} C_Q(a)$.

D.5 Sequence models and optimization

Each row is converted into a left-padded, within-trajectory causal window. A padding indicator is appended to every token, taking value one on real observations and zero on left padding. The synthetic window length is 14; the Rossmann window length is 21. Feature means and standard deviations are estimated from training windows only and then applied unchanged to validation and test windows.

The main encoder is a one-layer GRU. Its last hidden output is passed through a linear softmax head. The current action is part of the final token processed by the GRU. The diagnostic MLP flattens the complete lag window and applies

$$\text{Linear}(Ld, m) \longrightarrow \text{GELU} \longrightarrow \text{Linear}(m, m) \longrightarrow \text{GELU} \longrightarrow \text{Linear}(m, K),$$

where L is the context length, d includes the padding indicator, m is the hidden width, and K is the number of output classes. No Transformer or dropout layer is used.

For the synthetic experiments, the inventory input dimension is 11, with 11 structural shock classes and 121 direct classes. The queue input dimension is 8, including the padding indicator, with 9 structural shock classes. The resulting parameter counts are 4,683 for the inventory structural GRU, 6,379 for the inventory structural MLP, 8,313 for the inventory direct GRU, 4,329 for the queue structural GRU, and 4,969 for the queue structural MLP. The direct queue GRU contains 4,560–4,659 parameters, depending on K_Q.

Training uses full-batch validation after every epoch. A checkpoint is retained when validation observed-continuation NLL improves by at least 10^{-5}; training stops after the applicable number of consecutive nonimproving epochs shown in Table 5. Hyperparameters and the stopping rule were fixed before test evaluation, and all 39 archived neural fits stopped before their applicable epoch caps. All models use deterministic CPU execution. For a synthetic DGP seed s, model seeds are s for the structural GRU, $s + 1000$ for the structural MLP, and $s + 2000$ for the direct GRU.

D.6 Synthetic splits and evaluation metrics

For each DGP seed, we generate 900 independent 30-period trajectories. Assignment is by entire trajectory, in generated episode order:

	train	validation	test
episodes	540	180	180
decision epochs	16,200	5,400	5,400.

No trajectory contributes prefixes to more than one split. The five complete DGP/training seeds are

$$11, \quad 23, \quad 37, \quad 53, \quad 71.$$

For a held-out context j, let $p_j^\star, \widehat{p}_j$ denote the true and learned shock laws and let $q_j^\star, \widehat{q}_j$ be their

Table 5: Frozen neural-training protocol.

Setting	Structural synthetic	Direct synthetic	Rossmann system
Context length	14	14	21
Hidden width	32	32	40
Optimizer	AdamW	AdamW	AdamW
Learning rate	10^{-3}	10^{-3}	10^{-3}
Weight decay	10^{-4}	10^{-4}	10^{-4}
Batch size	512	512	512
Gradient-norm cap	1	1	1
Maximum epochs	70	400	80
Early-stopping patience	10	30	12
CPU threads	8	8	8

pushforwards. We report

$$\mathrm{ShockKL} = \frac{1}{N}\sum_{j=1}^{N} \mathrm{KL}(p_j^\star \| \widehat{p}_j), \tag{47}$$

$$\mathrm{ObservableKL} = \frac{1}{N}\sum_{j=1}^{N} \mathrm{KL}(q_j^\star \| \widehat{q}_j), \tag{48}$$

$$\mathrm{InvalidMass} = \frac{1}{N}\sum_{j=1}^{N} \left[1 - \sum_{z \in \mathcal{S}(H_j, A_j)} \widehat{q}_j(z)\right]. \tag{49}$$

Structural invalid mass is zero by construction. The observed NLL is (36) for structural models and the ordinary categorical NLL of the realized observable continuation for direct models.

Synthetic decision regret is computed on the first 1,200 held-out decision epochs in deterministic episode order. This computational subsample and its ordering are common to all models. The metric is one-step counterfactual regret at held-out states, not the cost of recursively deploying the learned policy over a new closed-loop trajectory.

Data-processing diagnostic. To isolate the reactor mechanism independently of neural estimation, we exponentially tilt every held-out true shock law:

$$p_{j,\delta}(x) = \frac{p_j^\star(x) e^{\delta x}}{\sum_v p_j^\star(v) e^{\delta v}}, \qquad \delta \in \{0.10, 0.20, 0.40\}. \tag{50}$$

For inventory, x is demand; for the queue, x is net input. The code enumerates both distributions and their pushforwards and checks pointwise that

$$\mathrm{KL}\big(\Phi_\# p_j^\star \,\big\|\, \Phi_\# p_{j,\delta}\big) \le \mathrm{KL}(p_j^\star \| p_{j,\delta}).$$

We also report the observable-to-shock KL ratio. Learned-model contraction ratios are stratified by available inventory 0–2, 3–5, and 6+, and by queue length 0, 1–2, and 3+. Ratios with shock KL below 10^{-10} are omitted.

Joint-KL chain-rule diagnostic. For each structural model, the cumulative conditional-KL estimate through period h is

$$K_h = \sum_{t=1}^{h} \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathrm{KL}(q^\star_{it} \,\|\, \widehat{q}_{it})\,.$$

The corresponding shock-KL upper curve replaces q by p. On the realized held-out logging trajectories, we independently compute the path likelihood-ratio estimate

$$\widehat{L}_h = \sum_{t=1}^{h} \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \log \frac{q^\star_{it}(Z_{i,t+1})}{\widehat{q}_{it}(Z_{i,t+1})}.$$

To audit agreement, define the episode-level centered difference

$$D_i = \sum_{t=1}^{30} \left[\log \frac{q^\star_{it}(Z_{i,t+1})}{\widehat{q}_{it}(Z_{i,t+1})} - \mathrm{KL}(q^\star_{it} \| \widehat{q}_{it}) \right],$$

and report

$$Z_{\text{chain}} = \frac{|\overline{D}|}{s_D/\sqrt{180}}.$$

The independent sampling unit in this diagnostic is the complete test episode, not an individual transition.

D.7 Persistent-demand pricing

At time zero, the manager commits to a price p for H periods. One market state Y is drawn and remains fixed throughout the trajectory:

$Y \sim \text{Bernoulli}(1/2)$ under the true law, $\quad Y \sim \text{Bernoulli}(1/2+\delta)$ under the misspecified law.

Demand and per-period profit are

$$D_t(p) = 4 - p + Y, \qquad R_t(p) = pD_t(p).$$

Since Y is drawn once and then persists, the joint trajectory KL is independent of H:

$$\kappa_\delta = \mathrm{KL}(\text{Bern}(1/2) \,\|\, \text{Bern}(1/2+\delta)) = -\frac{1}{2}\log(1-4\delta^2). \tag{51}$$

True and misspecified expected total profits are

$$J^\star_H(p) = Hp(4.5 - p), \qquad \overline{J}_H(p) = Hp(4.5 + \delta - p).$$

For continuous prices $p \in [1,3]$, $J^\star_H$ is $2H$-strongly concave. The true and plug-in prices are

$$p^\star = 2.25, \qquad \widehat{p} = 2.25 + \frac{\delta}{2},$$

and true regret is

$$J^\star_H(p^\star) - J^\star_H(\widehat{p}) = \frac{H\delta^2}{4} = \Theta(H\kappa_\delta). \tag{52}$$

Equivalently, negative profit is $2H$-strongly convex. This is strong convexity in the operational decision p, not in the parameter used to fit the Bernoulli law. The latter is neither assumed nor used.

To isolate the discontinuous policy-switching geometry, set the price separation to $d = 1$ and restrict the action set to

$$p_{\text{low}} = 2.25 + \frac{\delta}{4} - \frac{d}{2}, \qquad p_{\text{high}} = 2.25 + \frac{\delta}{4} + \frac{d}{2}.$$

The true law selects p_{low}, whereas the misspecified law selects p_{high}. Therefore

$$J_H^\star(p_{\text{low}}) - J_H^\star(p_{\text{high}}) = \frac{Hd\delta}{2} = \Theta(H\sqrt{\kappa_\delta}). \tag{53}$$

This discrete instance is not strongly concave over its action set. Together, (52) and (53) separate horizon-sensitive transfer from the stabilizing effect of downstream curvature. More precisely, the same continuation-law error and the same ambient quadratic profit appear in both cases. What fails in the two-price case is the convex decision domain and stable first-order optimizer map; ambient curvature alone does not prevent a discontinuous policy switch.

We use

$$H \in \{4, 8, 16, 32, 64, 128\}$$

and

$$\delta \in \{2^{-8}, 2^{-7}, 2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 0.20, 0.25\}.$$

The displayed log–log slopes are computed at $H = 4$ using $\delta \le 2^{-4}$. The formulas are also checked using 200,000 common-random-number trajectories at $H = 64$, $\delta = 0.08$, and seed 123. Trajectories, rather than repeated periods within a trajectory, are the Monte Carlo sampling units.

D.8 Semi-synthetic Rossmann inventory

Interpretation and data preparation. Rossmann daily `Sales` is aggregate turnover, not observed SKU-level demand. We therefore use it only as a demand proxy supplying real serial, promotional, and calendar variation. The construction supplies realistic path variation but does not identify causal effects or monetary savings at Rossmann; the precise one-step estimand is defined below.

The source archive is the public Rossmann data mirror at

`https://files.fast.ai/part2/lesson14/rossmann.tgz`,

whose verified SHA-256 digest is

`24ccc6757ea620069716482ffec2d4e17689d54a5d7acf27f68adb02069c00aa`.

Only `Store`, `DayOfWeek`, `Date`, `Sales`, `Open`, `Promo`, `StateHoliday`, and `SchoolHoliday` are loaded. Contemporaneous `Customers` is excluded.

For store s, let m_s be its median positive sales on open days in the training period. The latent demand proxy is

$$D_{s,t}^R = \begin{cases} \left[\text{round}\left(6\,\text{Sales}_{s,t}/m_s\right)\right]_0^{20}, & \text{Open}_{s,t} = 1, \\ 0, & \text{Open}_{s,t} = 0. \end{cases} \tag{54}$$

The medians m_s are estimated using training dates only. The observed cap rate is 2.7203×10^{-5}, or approximately 0.0027%.

Paper mode selects the first 40 store identifiers with complete daily coverage through June 30, 2014, consulting only the training period when forming the cohort. The selected identifiers are $1, \dots, 40$; no validation- or test-period missingness criterion is applied. The fixed chronological split is

training:	2013-01-01 through 2014-06-30,	21,840 store-days,
validation:	2014-07-01 through 2014-12-31,	6,440 store-days,
test:	2015-01-01 through 2015-07-31,	8,480 store-days.

Synthetic inventory system. Inventory capacity is 12, the maximum order is 7, initial inventory is uniform on $\{3, \dots, 8\}$, and the lost-sales reactor is again (40), with D_{t+1} replaced by $D^R_{s,t}$. The sales EWMA is initialized at $M^R_{s,0} = 0.35$ and updated as

$$M^R_{s,t+1} = 0.90 M^R_{s,t} + 0.10 \, \frac{S_{s,t+1}}{20}.$$

The observable base-stock target is

$$T^R_{s,t} = \begin{cases} \left[\text{round}\{4 + 4M^R_{s,t} + 1.5P_{s,t}\}\right]_0^{12}, & \text{Open}_{s,t} = 1, \\ I_{s,t}, & \text{Open}_{s,t} = 0. \end{cases}$$

Let

$$u^R_{s,t} = \min\{7, 12 - I_{s,t}\}, \qquad b^R_{s,t} = [\, T^R_{s,t} - I_{s,t} \,]_0^{u^R_{s,t}},$$

and define the clipped neighborhood of $b^R_{s,t} - 1, b^R_{s,t}, b^R_{s,t} + 1$ as before. On open days, the logging rule uses 0.35 uniform exploration over all feasible orders and 0.65 probability on the clipped neighborhood. On closed days, uniform exploration is disabled and the action is drawn from that neighborhood. All inventory states, orders, censoring, and costs are therefore synthetic.

The feature token consists of

$$\begin{aligned} &(I_{s,t}, A_{s,t}, Y_{s,t}, S_{s,t}, B^I_{s,t}, \text{Open}_{s,t}, P_{s,t}, \text{SchoolHoliday}_{s,t}, \\ &\text{StateHolidayBinary}_{s,t}, \sin(2\pi \, \text{DayOfWeek}_{s,t}/7), \cos(2\pi \, \text{DayOfWeek}_{s,t}/7), \\ &\sin(2\pi \, \text{DayOfYear}_{s,t}/365.25), \cos(2\pi \, \text{DayOfYear}_{s,t}/365.25), M^R_{s,t}), \end{aligned}$$

followed by a 40-dimensional store indicator. Including the padding indicator gives input dimension 55.

The structural models output a 21-class categorical law for the hidden demand proxy and use the exact censored likelihood (36). The direct GRU predicts one of $13^2 = 169$ sales–next-inventory pairs. Parameter counts are 12,501 for the structural GRU, 48,741 for the structural MLP, and 18,569 for the direct GRU.

Semi-synthetic evaluation. The system path and randomized logging actions are fixed using seed 20260714. Neural initializations are $101, 103, 107$, with the same $+1000$ and $+2000$ offsets for the MLP and direct GRU. We report observed continuation NLL, invalid probability, and the realized one-step cost (42). Structural models additionally receive a latent-proxy NLL diagnostic using $D^R_{s,t}$, which is retained only for evaluation and is not exposed after censoring during training.

At every test store-day, the plug-in rule evaluates all feasible current orders, chooses the order with the smallest learned expected cost, and is scored on the realized demand proxy. The logging-policy benchmark uses the synthetic logged order. The perfect-information benchmark minimizes the same one-period realized cost after observing that day's demand proxy. These

comparisons use all 8,480 held-out store-days and condition on inventory states generated by the logging system. They do not recursively replace future states with states generated by the learned policy. Accordingly, the reported Rossmann cost is exclusively a one-step plug-in comparison on held-out logging states, not sequential deployment and not a field-savings estimate.

D.9 Paired statistical inference

For synthetic neural comparisons, the uncertainty unit is the complete DGP/training seed. If D_s is the metric difference between two models trained and evaluated on seed s, the reported 95% paired interval is

$$\overline{D} \pm t_{0.975,\,4}\frac{s_D}{\sqrt{5}}. \tag{55}$$

Individual transitions are not treated as independent observations.

For Rossmann, the three neural initializations are first averaged within each store, model, and metric. The 40 stores are then used as the clusters in the same Student-t construction:

$$\overline{D} \pm t_{0.975,\,39}\frac{s_D}{\sqrt{40}}. \tag{56}$$

All model comparisons are paired within seed or store. The Rossmann intervals quantify cross-store variation conditional on the realized semi-synthetic calendar and inventory construction; they are not causal field uncertainty.

Because the primary store intervals condition on the fitted three-seed ensemble, we report two additional finite-initialization sensitivities. First, we average each paired model difference across stores within each of the three seeds and apply a paired Student-t interval with two degrees of freedom. Second, we retain all seed–store paired differences and form a two-way CR1 sandwich variance with seed and store clustering, subtracting the intersection cluster contribution; its critical value is also $t_{0.975,2}$. With only three seed clusters these intervals are sensitivity analyses, not asymptotic precision claims. Section 5.3 reports their numerical values and keeps model-cost claims directional whenever the two-way interval contains zero.

Finite-class Monte Carlo uncertainty is reported using the replication-level standard error $s/\sqrt{1000}$. Pricing curves use exact formulas; the separate common-random-number simulation reports trajectory-level Monte Carlo standard errors.

D.10 Additional trajectory and Rossmann diagnostics

See Figure 6 below.

D.11 Reproducibility audit

The archived paper run used Python 3.12.13, NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, scikit-learn 1.8.0, PyTorch 2.6.0 + cpu, Matplotlib 3.10.8, and seaborn 0.13.2. PyTorch deterministic algorithms were enabled, data-loader shuffling used a seeded generator, and no worker subprocesses were used. The complete paper protocol ran on CPU in 1,525.9 seconds.

The replication package contains separate tests for reactor balance, censoring-event masks, pushforward normalization, data processing, and pricing rates. The frozen run passed the following audit gates:

1. every planned synthetic and Rossmann seed completed and every archived metric was finite;

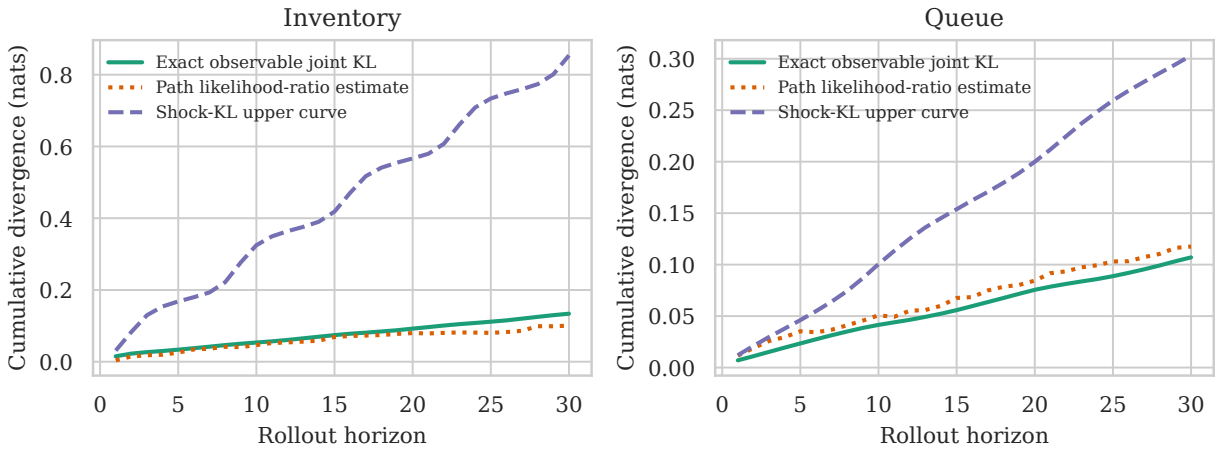

Figure 5: Joint-KL accumulation in inventory and queueing. Exact conditional observable KL is accumulated by the autoregressive chain rule. The shock curve is its data-processing upper bound, and the path likelihood-ratio curve is an independent-trajectory Monte Carlo estimate. Curves are averaged over five seeds.

2. all structural continuation laws assigned probability one to their context-specific feasible support;
3. the largest pointwise value of $\mathrm{KL}(\Phi_{\#}p^{\star}\|\Phi_{\#}p) - \mathrm{KL}(p^{\star}\|p)$ was 8.33×10^{-17}, consistent with floating-point error;
4. structural invalid mass was exactly zero in every synthetic and Rossmann result row;
5. every primary structural-GRU chain-rule diagnostic satisfied $Z_{\mathrm{chain}} < 1.92$;
6. one nonprimary inventory-MLP seed had $Z_{\mathrm{chain}} = 2.29$; it was retained rather than filtered;
7. the pricing log–log slopes were 0.499 for the discrete near-tie and 0.999 for the strongly concave problem;
8. the finite-class normalized quantity $n\,\mathrm{KL}/H$ lay between 0.472 and 0.503 over the complete grid; and
9. the Rossmann file digest, store identifiers, dates, split sizes, and demand-proxy cap rate matched the frozen audit record.

Raw outputs are stored in tidy CSV form, with one row identifying the experiment, seed, model, metric, and any horizon, stratum, or perturbation level. Tables and figures are regenerated from these frozen files without retraining. The run manifest records the complete configuration, software versions, platform, CPU thread count, and elapsed time.

Interpretive boundaries. Exact KL and data-processing statements are reported only for the synthetic systems, where the true shock law is known. Projection is a support-repair baseline and should not be interpreted as learning the structural shock law. The pricing and finite-class studies are controlled mathematical diagnostics. Finally, Rossmann supplies realistic sales-path variation but neither true uncensored demand nor randomized field interventions; its estimand remains the one-step comparison defined in the semi-synthetic evaluation protocol.

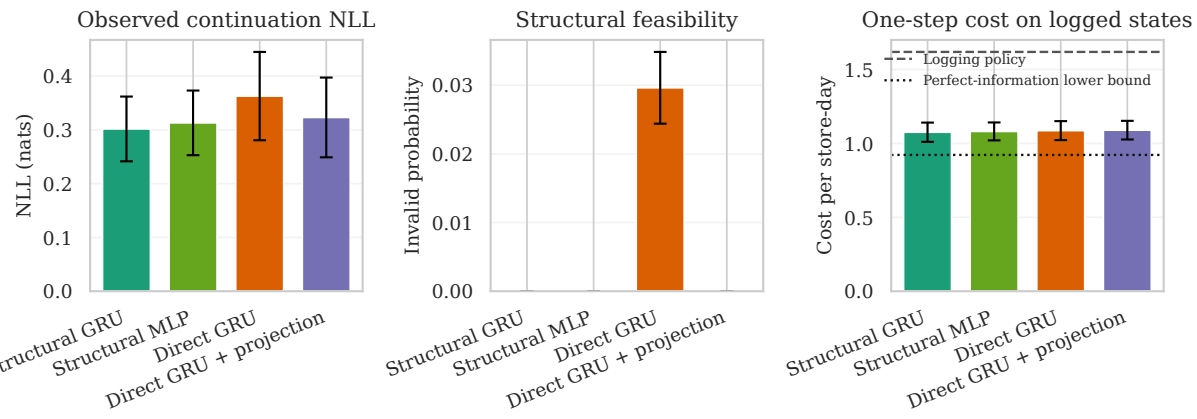

Figure 6: Semi-synthetic Rossmann inventory diagnostics: observed-continuation NLL, invalid probability, and constructed one-step cost. Dashed and dotted lines in the cost panel show the synthetic logging rule and one-step perfect-information lower bound.

References

Dimitris Bertsimas and Nathan Kallus (2020). From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch (2021). Decision Transformer: Reinforcement learning via sequence modeling. *Advances in Neural Information Processing Systems*, 34.

Priya L. Donti, Brandon Amos, and J. Zico Kolter (2017). Task-based end-to-end model learning in stochastic optimization. *Advances in Neural Information Processing Systems*, 30.

Adam N. Elmachtoub and Paul Grigas (2022). Smart "Predict, then Optimize." *Management Science*, 68(1):9–26.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos (2010). Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems*, 23.

Dylan J. Foster, Adam Block, and Dipendra Misra (2024). Is behavior cloning all you need? Understanding horizon in imitation learning. *Advances in Neural Information Processing Systems*, 37:120602–120666.

David Ha and Jürgen Schmidhuber (2018). Recurrent world models facilitate policy evolution. *Advances in Neural Information Processing Systems*, 31.

Elad Hazan, Shai Shalev-Shwartz, and Nathan Srebro (2025). Research program: Theory of learning in dynamical systems. *arXiv preprint arXiv:2512.19410*.

Hui Jiang (2023). A latent space theory for emergent abilities in large language models. *arXiv preprint arXiv:2304.09960*.

Kaggle (2015). Rossmann Store Sales: Forecast sales using store, promotion, and competitor data. https://www.kaggle.com/c/rossmann-store-sales.

Saman Lagzi, Ningyuan Chen, and Joseph Milner (2026). Using neural networks to guide data-driven operational decisions. *Management Science*, published online June 15, 2026. doi:10.1287/mnsc.2023.04141.

Dennis Victor Lindley (1952). The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(2):277–289.

Sam Lobel and Ronald Parr (2024). An optimal tightness bound for the simulation lemma. *arXiv preprint arXiv:2406.16249*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). Improving language understanding by generative pre-training. OpenAI technical report.

Dhruv Rohatgi, Adam Block, Audrey Huang, Akshay Krishnamurthy, and Dylan J. Foster (2025). Computational-statistical tradeoffs at the next-token prediction barrier: Autoregressive and imitation learning under misspecification (extended abstract). *Proceedings of the 38th Conference on Learning Theory*, PMLR 291:4831–4837. Full version: *arXiv preprint arXiv:2502.12465*.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191.

Simons Institute for the Theory of Computing (2020). Mathematics of online decision making. Workshop in the Theory of Reinforcement Learning program, October 26–30, 2020. https://simons.berkeley.edu/workshops/mathematics-online-decision-making.

Hanzhao Wang, Guanting Chen, Kalyan Talluri, and Xiaocheng Li (2025). OMGPT: A sequence modeling framework for data-driven operational decision making. *arXiv preprint arXiv:2505.13580*.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang (2023). Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 36:15614–15638.

Noam Wies, Yoav Levine, and Amnon Shashua (2023). The learnability of in-context learning. *Advances in Neural Information Processing Systems*, 36:36637–36651.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma (2022). An explanation of in-context learning as implicit Bayesian inference. *International Conference on Learning Representations*.

Yunbei Xu, Yuzhe Yuan, and Ruohan Zhan (2026). Autoregressive learning in joint KL: Sharp oracle bounds and lower bounds. *arXiv preprint arXiv:2605.12316*. doi:10.48550/arXiv.2605.12316.

Oğuz Kaan Yüksel and Nicolas Flammarion (2025). Generalization bounds for autoregressive processes and in-context learning. *NeurIPS 2025 Workshop on Principles of Generative Modeling*.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang (2023). What and how does in-context learning learn? Bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*.