MEMORY FOR LATE EVIDENCE: A DELAY–RESOLUTION PHASE LAW

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequential state estimators are designed to forget: once a compact state is sufficient for predicting the future, the observations that produced it can be discarded. Late evidence exposes a fundamental limitation of this principle. A state may be sufficient for forward prediction yet insufficient to revise the present when an observation about a discarded timestamp arrives later. We characterize the minimum persistent memory required to support such revisions under arbitrary legal arrival orders. For a class of positive, invertible K-state models with bounded, locally rich on-time evidence and retrodictively observable late evidence, the optimal deterministic code satisfies

$$B_\epsilon^\star(D) = \Theta((K-1)\min\{D+1, \log(1/\epsilon)\}\log(1/\epsilon))\,.$$

Thus, correction-state memory is delay-limited only up to the point at which the required resolution dominates: beyond this horizon, memory becomes resolution-limited and no longer grows with physical delay. The upper bound stores a quantized cut belief together with the resolvable suffix, while a matching response-packing converse shows that every retained direction is necessary, ruling out an explanation based solely on truncating a mixing filter. Without bounded evidence, even delay one can preclude any finite uniformly accurate code. These results identify statistical resolution, rather than physical delay, as the relevant capacity variable for correctable state beyond the mixing horizon. A packed cut–suffix construction realizes the predicted memory and replay-work saturation, while a controlled recurrent bottleneck diagnostic reveals the forward–revision tension on prediction-equivalent histories. Overall, the theory provides a concrete criterion for sequential representations under late evidence: preserve the responses required for future corrections, not merely the information needed for prediction from the current state.

1 INTRODUCTION

A Markov filter is designed to forget. Once the current belief b_T has been computed, the observations that produced it can be discarded without affecting ordinary forward prediction. Late evidence changes this principle. If an observation about time $\tau < T$ arrives after the intervening history has been compressed, revising b_T requires information about how X_T remained coupled to X_τ. The marginal law of X_T alone need not contain this information.

This issue arises whenever timestamped evidence can arrive after the state has already been updated, including asynchronous sensing, delayed multimodal streams, and latent-state filtering. More generally, sequential ML often treats two histories as equivalent once they induce the same predictive state. Late evidence requires a stronger notion of equivalence: two histories must also produce the same response to future observations targeting discarded timestamps. Thus, a state may be sufficient for prediction while being insufficient for revision.

Figure 1 illustrates the distinction with the smallest nontrivial example. Two joint posteriors on (X_{T-1}, X_T) can have the same current marginal $\Pr(X_T = 0) = 0.5$, while conditioning on the same late observation $X_{T-1} = 0$ yields revised probabilities 0.9 and 0.1. Before the late observation, the two histories are indistinguishable to a forward predictor; afterward, they require different corrections. The current belief is therefore prediction-sufficient but not *revision-sufficient*.

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

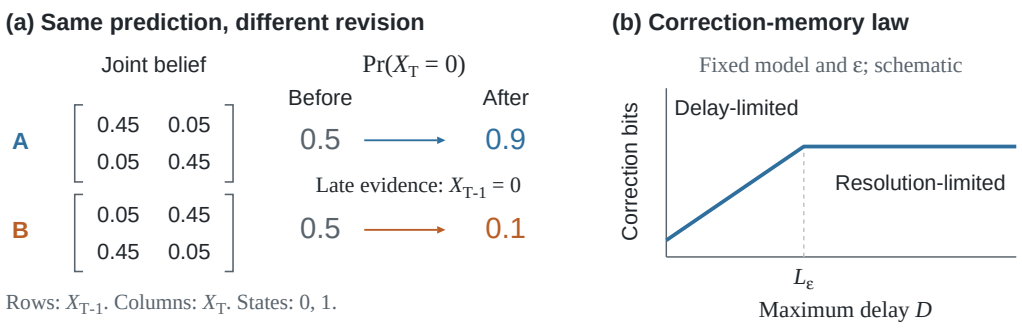

Figure 1: **Late evidence changes the relevant notion of memory.** *(a)* Two joint beliefs share $\Pr(X_T = 0) = 0.5$, but conditioning on $X_{T-1} = 0$ yields 0.9 and 0.1. *(b)* The correction-memory law at fixed model and ϵ: memory grows with delay until the resolvable horizon L_ϵ, then saturates.

Out-of-sequence estimation already studies delayed measurements, including storage-efficient and selective particle filters (Orguner & Gustafsson, 2008; Berntorp et al., 2012; Oreshkin et al., 2011). Short-memory prediction similarly studies when distant history can be discarded without substantially affecting forward prediction (Sharan et al., 2018; Han et al., 2024). These approaches, however, do not characterize the minimum persistent state required to remain correct after arbitrary legal insertions of future evidence. We isolate this question in a known finite-state model, separating the fundamental memory requirement from the implementation details of a particular filtering or learning algorithm.

What this paper establishes

Problem. We seek the minimum number of persistent deterministic bits required to process timestamped likelihood factors in arbitrary legal subset and order, while decoding the terminal belief within ϵ total variation after *every* arrival prefix.

Memory law. For $q = K - 1$ and $x = \log(1/\epsilon)$, under positive invertible transitions, bounded locally rich on-time evidence, and a checkable response-observability condition,

$$B^\star_\epsilon(D) = \Theta(q \, \min\{D + 1, x\} \, x) \, .$$

Interpretation. Correction memory is initially *delay-limited*, $\Theta(q(D + 1)x)$, but becomes *resolution-limited*, $\Theta(qx^2)$, once the delay exceeds the resolvable horizon L_ϵ. Thus, physical delay can continue to grow while the information that remains relevant for correction does not.

Tightness. A cut–suffix construction achieves the upper bound, while a matching response-packing converse shows that the retained revision directions are necessary. The result is therefore a memory characterization rather than merely a truncation heuristic.

The proof has two complementary parts. The upper bound uses contraction to retain a quantized cut belief together with only the resolvable suffix. The lower bound constructs a packing of response directions induced by legal late observations: before mixing erases a direction, different histories remain distinguishable by a common future insertion. Consequently, the same resolution horizon controls both the sufficient and necessary memory scales. The construction also yields a replay procedure whose work saturates once the resolvable suffix has been reached, although the theorem's optimality claim concerns persistent memory rather than replay time.

The resulting criterion is simple: a sequential representation is safe for late evidence only if it preserves the part of its history that future corrections can still resolve. It need not retain the entire physical delay window once older responses fall below the target resolution. In this sense, the relevant capacity of a correctable state is determined not by delay alone, but by the interaction between delay, state dimension, and statistical resolution.

The assumptions are essential. Without bounded evidence, even delay one can preclude any finite uniformly accurate deterministic code. The theory therefore isolates the finite-memory regime in which

late evidence remains locally observable while sufficiently old responses become indistinguishable at resolution ϵ.

2 Related Work

Table 1: **Related frameworks optimize different objects.** Only the last row measures persistent bits under recursive late-evidence insertion.

Framework	Arrival protocol	Guarantee	Optimized quantity
Predictive / belief abstractions	in order	prediction or value accuracy	state representation
Out-of-sequence filters	timestamped late data	tracking accuracy, sometimes after selection	implementation storage and work
Causal rate–distortion	sequential coding	expected distortion	information rate
Delayed-observation RL	delayed stream	closed-loop return	policy
This paper	arbitrary subset and order	pointwise TV $\leq \epsilon$ at every prefix	stored bits $B^\star_\epsilon(D)$

Memory for prediction and control. Short-memory predictors and structured POMDP learners can sometimes replace full history by a recent window (Sharan et al., 2018; Han et al., 2024; Efroni et al., 2022). Predictive states and approximate information states preserve future predictions or values (Littman et al., 2001; Subramanian et al., 2022; Kara & Yüksel, 2022; Kim, 2026). Belief-compression and observable-POMDP methods instead seek tractable belief or decision-relevant subspaces (Boyen & Koller, 1998; Poupart & Boutilier, 2002; Roy & Gordon, 2002; Golowich et al., 2022). These results process evidence chronologically. We ask what a summary must retain when evidence later targets a discarded timestamp.

Delayed and out-of-sequence observations. Out-of-sequence filters retain or reconstruct past states (Evans & Krishnamurthy, 1999; Mallick et al., 2002; Bar-Shalom, 2002; Zhang & Bar-Shalom, 2012). Storage-efficient particle filters use auxiliary smoothers or reduced retained statistics (Orguner & Gustafsson, 2008; Berntorp et al., 2012), while selective schemes discard low-utility measurements (Oreshkin et al., 2011). These works already reduce implementation storage or work; our novelty is not low-memory OOSM filtering. They optimize tracking accuracy under algorithm-specific approximation or selection. We instead accept every legal insertion and characterize minimax persistent bits pointwise after every prefix. Fixed-lag smoothers and stability analyses make the window explicit or justify ignoring old data (Shue & Dey, 2002; Houssineau et al., 2019; Shue et al., 1998; Le Gland & Oudjane, 2004; van Handel, 2008; Douc et al., 2009; Karjalainen et al., 2023). Karamzade et al. process random out-of-order observations through sequential belief updates in model-based RL (Karamzade et al., 2026); we condition on a known model and ask the orthogonal minimum-state question under every legal prefix. Other delayed-observation RL work targets policy performance (Kim & Jeong, 1987; Bouteiller et al., 2021). Retroactive data structures also update past operations, but preserve program semantics rather than Bayesian responses (Demaine et al., 2007). We turn contraction into a recursive-state upper bound and add the matching response-packing converse that forgetting alone does not provide.

The closest results impose a different accuracy requirement. OOSM and fixed-lag methods bound tracking error for an algorithm that retains, reconstructs, or selects past information, while approximate-sufficiency work codes a statistic for a static experiment (Orguner & Gustafsson, 2008; Oreshkin et al., 2011; Hayashi & Tan, 2018). Neither asks one deterministic recursive code to remain pointwise accurate after every prefix of every legal timestamped insertion word. Consequently, their guarantees do not imply the response-metric converse or the persistent-bit law in Theorem 3.

State equivalence and lossy representations. Prediction and retrodiction induce distinct causal states (Ellison et al., 2009; Shalizi & Crutchfield, 2001); Blackwell equivalence, Myhill–Nerode quotients, and approximate sufficiency formalize history merging (Blackwell, 1953; Nixon, 2026; Hayashi & Tan, 2018). HMM observability and minimal realization instead ask when latent dynamics are distinguishable (van Handel, 2009; Huang et al., 2016). Rate–distortion and rate-constrained

control use average or communication-limited criteria (Marzen & Crutchfield, 2014; Charalambous & Stavrou, 2013; Tatikonda & Mitter, 2004). Our state is deterministic, pointwise over sensor support, and accurate after every prefix. The contribution is the matched metric entropy of this timestamped-insertion class, complementing exact nonexistence results (Chaleyat-Maurel & Michel, 1984). Table 1 aligns the neighboring problems by their arrival protocol, guarantee, and optimized quantity.

3 RETRODICTIVE MEMORY

Fix decision time T and maximum delay D. A known finite-state Markov model induces a positive posterior π on $X_{T-D:T}$. A late event $e = (\tau, v)$ multiplies the path density by a known local likelihood at $\tau < T$; write $U_e\pi$ for the normalized update. Legal event words form a prefix-closed family $\mathcal{W}_D$, and the required output is the terminal marginal $b(\pi) = \pi(X_T \in \cdot)$.

Definition 1 (Exact recursive state)**.** An exact state comprises an encoder C, event updates δ_e, and decoder d satisfying $C(U_e\pi) = \delta_e(C(\pi))$ and $d(C(\pi)) = b(\pi)$ on every reachable posterior and legal event. A Euclidean-valued C must be continuous; its minimum ambient dimension is the *continuous exact dimension*.

Continuity is essential: otherwise an open high-dimensional set can be hidden in one real number. This dimension is a geometric resource, not a floating-point bit count.

Definition 2 (Prefix-uniform stored bits)**.** A deterministic B-bit state has at most 2^B values, a deterministic update for each timestamped event, and a decoded terminal distribution. It is ϵ-accurate if

$$\sup_{\pi, w \in \mathcal{W}_D} \sup_{0 \le j \le |w|} \mathrm{TV}\Big(b(U_{w_{1:j}}\pi), \widehat{b}(\delta_{w_{1:j}}(C(\pi)))\Big) \le \epsilon.$$

Let $B^\star_\epsilon(D)$ be the minimum B.

Accounting. $B^\star_\epsilon(D)$ charges only persistent correction state. Throughout Theorems 2–3, the transport layer maintains admissibility externally and supplies only legal packets; the transition and sensor laws, update program, clock, and timestamp/value of the packet currently being processed are also external. If the filter must instead recover the exact pending mask, Theorem 4 gives an additional $\Theta(D)$ term. Transient arithmetic workspace is likewise not part of the stored code. The accuracy criterion is deterministic and pointwise over the declared observation support: it averages over neither histories nor arrival orders. Requiring accuracy after every prefix rules out a decoder that waits for an entire bundle before becoming correct.

Responses to future insertions. Under the external-scheduler interface above, two histories are retrodictively equivalent when every continuation legal for both histories has the same terminal output. This common-legal right congruence is the coarsest exact correction state. If admissibility must be reconstructed internally, equivalence additionally requires the same admissible continuations; the pending-mask cost is then charged as stated above. Related quotient constructions occur in statistical experiments, computational mechanics, predictive representations, and bounded-interaction abstraction (Blackwell, 1953; Shalizi & Crutchfield, 2001; Littman et al., 2001; Nixon, 2026). Our concern is its quantitative geometry under late updates.

Define the response metric

$$d_R(\pi, \pi') = \sup_{w \text{ common legal}} \mathrm{TV}\big(b(U_w\pi), b(U_w\pi')\big).$$

Proposition 1 (Response packing)**.** *Every prefix-uniform B-bit state satisfies*

$$B \ge \log_2 \mathcal{P}(\Pi, d_R, 2\epsilon),$$

where $\mathcal{P}$ is the strict packing number. The statement also holds for any declared restricted word family.

Proof. Suppose two histories π, π' share one code value. For every common legal continuation w, deterministic event updates leave them in the same code state and hence produce one decoded output

$\widehat{b}_w$. Prefix-uniform accuracy applies at the prefix ending in w, so

$$\mathrm{TV}(b(U_w\pi), b(U_w\pi')) \le \mathrm{TV}(b(U_w\pi), \widehat{b}_w) + \mathrm{TV}(\widehat{b}_w, b(U_w\pi')) \le 2\epsilon.$$

Thus two elements of a strict 2ϵ response packing cannot share a state. A B-bit state has at most 2^B values, giving the bound. Restricting the supremum to a declared subfamily of continuations uses the same argument. □

A metric cover alone is not enough for achievability: its cells need not stay closed after another delayed observation arrives. The upper construction in Section 4 is designed to remain closed under every legal update.

4 When Finite Memory Is Possible

The late-event response can be either arbitrarily brittle or uniformly compressible. The dividing assumptions are bounded evidence and conditional contraction, not delay alone.

What the assumptions do. Bounded likelihood fields prevent a rare packet from amplifying an arbitrarily small hidden distinction; Theorem 1 shows that some such restriction is necessary. Uniform positivity supplies the conditional contraction used by the constructive upper bound. Response observability is not needed for compression: it is the nondegeneracy condition that makes the matching lower bound possible. Transition invertibility enters that converse and the exact-dimension statement. The assumptions therefore separate an upper-bound mechanism, a lower-bound mechanism, and a necessity boundary rather than serving as one undifferentiated proof convenience.

4.1 A delay-one impossibility with unbounded evidence

Let $X \in \{0,1\}$ be uniform and let the terminal state Z be a noisy copy, $\Pr(Z \neq X \mid X) = \beta$ for $0 < \beta < 1/2$. Two conditionally independent sensors Y, Y' observe X through the same countable positive channel. For $k \in \mathbb{Z}$, choose $v_k = e^{-\lambda|k|}$ with $\lambda > c/2$ and

$$p(Y = k \mid X = 0) = v_k e^{kc/2}/A, \qquad p(Y = k \mid X = 1) = v_k e^{-kc/2}/A,$$

where symmetry gives a common finite normalizer A. Outcome k has likelihood ratio e^{kc}, so the ratios are finite pointwise but unbounded over the alphabet.

Theorem 1 (Rare late evidence)**.** *For this fixed delay-one HMM, the response metric has an infinite δ-packing with $\delta = (1-2\beta)\tanh(c/2) > 0$. No finite code is uniformly accurate for $\epsilon < \delta/2$.*

Proof sketch. After the on-time outcome $Y = 2n$, the log odds of $X = 0$ equal $2nc$. For two distinct histories n, m, the common legal late outcome $Y' = -(n+m)$ changes their log odds to $(n-m)c$ and $(m-n)c$. Their terminal probabilities differ by at least δ. Proposition 1 yields the claim. The separating outcome always has positive probability, although that probability vanishes along the packing. □

This is a fixed-model form of conditioning amplification, consistent with broader brittleness phenomena in Bayesian inference (Owhadi et al., 2015). It prevents any theorem based only on an ergodic transition.

4.2 A finite-state constructive upper bound

Now consider a known K-state, possibly time-varying Markov chain with

$$0 < m < M < 1, \qquad m \le P_i(x, x') \le M.$$

The initial reduced logits are bounded. On-time local reduced log-likelihood fields lie in $[-H, H]^{K-1}$, and at most one delayed message per past timestamp adds a field in $[-A, A]^{K-1}$. Every subset and permutation is legal.

For any positive future weight g, the evidence-conditioned edge is

$$K_{i,g}(x, x') = \frac{P_i(x, x')g(x')}{\sum_u P_i(x, u)g(u)}.$$

It obeys

$$K_{i,g}(x, x') \geq \frac{m}{M} \frac{g(x')}{\sum_u g(u)}.$$

Thus its row total-variation diameter is at most $\kappa = 1 - m/M < 1$. Products contract by κ^L, even after arbitrary positive evidence on the suffix, because that evidence is absorbed into the backward weights. This is a direct Doob-minorization argument, compatible with classical filter-stability and coupling analyses (Le Gland & Oudjane, 2004; Douc et al., 2009; van Handel, 2008).

Theorem 2 (Finite-state upper bound)**.** *Let*

$$L_\epsilon = \left\lceil \frac{\log(2/\epsilon)}{-\log \kappa} \right\rceil, \qquad n_\epsilon(D) = \min\{D + 1, L_\epsilon + 1\}.$$

For every $D \geq 0$ and all sufficiently small ϵ, a recursively updateable deterministic state has prefix-uniform terminal error at most ϵ and stored code length

$$B_\epsilon(D) \leq C(K - 1)n_\epsilon(D) \log(1/\epsilon).$$

The constant depends on m, M, H, A and the initial-logit bound, but not on D.

Proof sketch. There are two branches. If $D < L_\epsilon$, retain every one of the $D + 1$ field blocks. Projection followed by rounding at spacing $\Delta = 2\epsilon/(D + 1)$ costs at most ϵ by a normalized-tilt telescope. The fixed-length grid index costs $O((K - 1)(D + 1)[\log(1/\epsilon) + \log(D + 1)])$ bits. In this branch $D + 1 = O(\log(1/\epsilon))$, so the bracket is $O(\log(1/\epsilon))$.

If $D \geq L_\epsilon$, ignore late messages older than L_ϵ. Full and truncated models share the retained suffix, so arbitrary disagreement at the left cut contracts to at most $\kappa^{L_\epsilon} \leq \epsilon/2$. Prediction bounds the cut logits by $\log(M/m)$, and the evidence assumptions put every retained field in a fixed box $[-\Lambda, \Lambda]^{K-1}$. Round the retained coordinates at spacing $\epsilon/(L_\epsilon + 1)$ on initialization and after their one possible late update. The same telescope costs at most $\epsilon/2$. Thus this branch uses $O((K - 1)(L_\epsilon + 1) \log(1/\epsilon))$ bits. Both constructions are closed under every legal event update; taking the appropriate branch proves the claim. □

Corollary 1 (Same-resource window comparison)**.** *Under the same fixed-width grid accounting, the explicit approximate code that retains all $D + 1$ field blocks uses*

$$B_{\text{full-grid}} = \Theta(q(D + 1)[\log(D + 1) + \log(1/\epsilon)])$$

bits, while the cut–suffix construction uses $O(q \min\{D + 1, \log(1/\epsilon)\} \log(1/\epsilon))$ bits. This compares two approximate recursive codes; it is not an exact-bit lower bound for a full-window representation.

The bound already displays the two regimes of the phase diagram. Before the mixing horizon, the code pays for all $D + 1$ timestamps. Afterwards it pays only for $L_\epsilon + 1 = \Theta(\log(1/\epsilon))$ timestamps. The remaining logarithm is the precision required per coordinate. Contraction-matched spacings $\Delta_d = \min\{2\Lambda, \epsilon/((L_\epsilon + 1)\kappa^d)\}$ reduce the leading constant of this explicit upper construction, but do not by themselves establish minimax optimality. The next section supplies matching lower bounds on the same all-delay regimes of an open response-observable class.

5 Why the Phase Law Is Unavoidable

The constructive upper bound says that old evidence may be forgotten after mixing. The result below answers the harder question: how much of the recent window must every recursive state preserve? Its response chart turns each still-resolvable timestamp into a distinguishable revision direction, giving the converse at the same horizon as the construction.

5.1 THE MATCHING RESPONSE LAW

Put $q = K - 1$. Let P be positive, invertible, and row stochastic, with stationary law π. In reduced-logit coordinates define $F(r) = \lambda(P^\top \sigma(r))$ and $r_\star = \lambda(\pi)$. Retain the bounded-field and update-legality hypotheses of Section 4: the entire on-time field support and the legal delayed-field set are bounded, initial logits range over a bounded set, and every subset and arrival permutation with at most one delayed message per timestamp is legal. In addition, the on-time field support contains a bounded open neighborhood of zero. The delayed-field set $\mathcal{V} \subset \mathbb{R}^q$ has interior. For $v \in \mathcal{V}$, put $A_v(r) = D[F(r+v) - F(r)]$.

The matrices $A_v(r)$ measure the part of a late update that survives prediction to the next state. Full rank means that legal late outcomes can reveal every local direction that the encoder might otherwise merge.

Lemma 1 (Automatic finite-probe observability)**.** *For every r, $\cap_{v\in\mathcal{V}} \ker A_v(r) = \{0\}$. At most q legal fields therefore give a full-column-rank stacked response, and one finite probe family works with positive margin on every compact base set.*

Indeed, if $DF(s)u$ were constant on the open set $r + \mathcal{V}$, analyticity would make F affine along u everywhere. Boundedness would force zero slope, contradicting nonsingularity of DF.

Theorem 3 (Field-rich matching law)**.** *For every fixed model above, continuous exact dimension is $q(D+1)$. There are model-dependent $c, C, \epsilon_0 > 0$ such that*

$$cq \min\{D+1, x\}x \le B^\star_\epsilon(D) \le Cq \min\{D+1, x\}x, \qquad x = \log(1/\epsilon),$$

for every $D \ge 0$ and $0 < \epsilon < \epsilon_0$.

Proof idea. Choose $n \le D+1$ recent timestamps and query how late outcomes at those timestamps change the terminal belief. Local full rank alone would give only local invertibility. Ordering the timestamps from newest to oldest instead makes the response map block triangular: after the newer blocks are recovered, each conditional diagonal map is co-Lipschitz on its controlled neighborhood. Backward substitution therefore gives global injectivity on the product cube, while products of nonsingular prediction Jacobians give $|\det D\Phi_n| \ge e^{-C_1 n^2}$. Change of variables then yields

$$\log \mathcal{P} \ge qnx - C_2 qn^2 - O(qn \log n).$$

Taking $n = \min\{D+1, \lfloor \gamma x \rfloor\}$ balances distinguishable volume against contraction and proves the lower bound through Proposition 1. A separated chart block is either the empty response or a displacement under one finite legal probe; the triangle inequality therefore selects a common legal continuation that separates the two terminal beliefs. The upper bound cuts the path at the same scale and stores a quantized cut belief together with the retained local fields. Because each coordinate is projected and rounded after its one possible late update, the finite grid remains closed under every legal insertion. Invariance of domain gives the separate exact-dimension claim. Appendix C contains the metric-transfer and chart details. This converse is the step that rules out interpreting the phase law as merely "mixing plus truncation": contraction supplies achievability, whereas the response volume proves that no recursive code can merge the surviving directions at resolution ϵ.

Corollary 2 (Open-neighborhood robustness)**.** *For every finite D, the response chart is injective with nonzero Jacobian on an open product neighborhood of on-time fields, and remains so under all sufficiently small parameter perturbations.*

The same conclusion does not require a continuum of *delayed* outcomes. The on-time channel still supplies the bounded open field neighborhood used to vary histories, while a finite delayed alphabet suffices whenever one positive outcome v obeys

$$\det[DF(r_\star + v) - DF(r_\star)] \ne 0. \tag{OC}$$

Proposition 2 (Finite delayed alphabet and genericity)**.** *Under* (OC)*, Theorem 3 holds. For each fixed K, positive full-rank transitions and positive binary late channels satisfying* (OC) *form an open dense, full-measure parameter subset.*

A quantitative version is available away from the singular set. Write $\mathcal{R}_K(P, \ell) = \det(P)\Psi_K(P, \ell)$ for the rational response certificate. After clearing its positive denominators, fix $\mathcal{R}_K = N_K/S_K$ with polynomial numerator N_K, positive polynomial denominator S_K, and $d_K = \deg N_K$. On a full-dimensional compact convex parameter body $\mathcal{C}$, let $\mu_\mathcal{C}$ be normalized Lebesgue measure.

Proposition 3 (Quantitative fixed-K margin)**.** *There is a finite constant $A_{\mathcal{C}}$, depending on $(\mathcal{C}, \mathcal{R}_K)$, such that*

$$\mu_{\mathcal{C}}\{|\mathcal{R}_K| \le \gamma\} \le \min\{1, A_{\mathcal{C}}\gamma^{1/d_K}\}.$$

When the on-time channel supplies one common bounded open field neighborhood, the matching-law constants are uniform on $\mathcal{C} \cap \{|\mathcal{R}_K| \ge \gamma\}$.

Appendix D proves both results and gives an explicit obstruction. The statement is quantitative at fixed K but is not uniform as the number of states grows.

5.2 Support: explicit nonscalar response charts

The binary anchor has transition correlation $0 < \rho < 1$, a bounded continuous on-time likelihood, and one delayed binary observation per past timestamp. Its retained-suffix chart satisfies

$$|\det D\Phi_L(0)| = 2^{-n}[q_0(1-\rho^2)]^L \rho^{L(L+1)/2}, \qquad n = L+1, \quad q_0 = \tanh^2 a.$$

It therefore has continuous exact dimension $D+1$ and $B^\star_\epsilon(D) = \Theta(\min\{D+1, x\}x)$.

This is not a scalar-monotonicity artifact. For the three-state reset model $P = \frac{1}{2}I + \frac{1}{6}\mathbf{1}\mathbf{1}^\top$,

$$DF(0) = \tfrac{1}{2}I, \qquad \det[DF((\log 2, 0)) - DF(0)] = -1/980.$$

The genuine 2×2 response blocks give continuous exact dimension $2(D+1)$ and the same matching approximate-bit law. Appendices E–F contain the scalar and block proofs.

5.3 Support: scaling with the number of states

The fixed-K theorem permits its conditioning margin to deteriorate with K. To make this dependence visible, set $\rho = 1/8$, $a = h = \log(4/3)$, and

$$P_K = \rho I + (1-\rho)\mathbf{1}\mathbf{1}^\top/K.$$

The bounded zero-sum softmax sensor and categorical delayed channel are specified in Appendix J. Prediction followed by any legal update has contraction coefficient $\kappa = 4/9$ for every K. For delayed outcome one, the response matrix has arrow form

$$A_K = \begin{pmatrix} \lambda_\parallel & c_K\mathbf{1}^\top \\ 0 & \lambda_\perp I_{K-2} \end{pmatrix}, \quad \lambda_\perp = -\frac{7}{8(24K+7)}, \quad c_K = -\frac{21K}{(24K+7)(25K+7)},$$

where $\lambda_\parallel = 7(24K^2 - 49K - 7)/[8(24K+7)(25K+7)]$. Its explicit inverse yields

$$\sigma_{\min}(A_K) \ge (200K)^{-1}. \tag{1}$$

Corollary 3 (Growing-state reset family)**.** *For this family, there exist absolute $c, C, C_0, \epsilon_0 > 0$, independent of K, D, ϵ, such that*

$$cq\min\{D+1, x\}x \le B^\star_\epsilon(D, K) \le Cq\min\{D+1, x\}x$$

whenever $0 < \epsilon < \epsilon_0$ and $x \ge C_0 \log K$. Continuous exact dimension is $q(D+1)$.

The lower chart pays $O(qn\log K)$ for the response margin in addition to $O(qn^2 + qn\log n)$. The condition $x \ge C_0 \log K$ absorbs that cost at $n = \min\{D+1, \Theta(x)\}$. Below this condition, the explicit packing inequality remains valid but the simplified uniform law is not asserted. The proof establishes the existence of an absolute C_0; it does not provide a useful numerical threshold. Section 6 therefore checks the response formulas rather than treating finite (K, ϵ) pairs as evidence for the asymptotic threshold.

5.4 Beyond a fixed decision time

Revealed inputs preserve the correction law, while a scheduler-exposed stream admits uniform-in-time correction with $O(T_{\mathrm{mv}}(K)L_\epsilon)$ materialization work. Storing the exact pending-packet mask adds $\Theta(D)$. Standard approximate-state transfer gives a passive decision bound (Subramanian et al., 2022; Kara & Yüksel, 2022; Kim, 2026), not a control theorem for action-dependent POMDPs. Appendix K states and proves these extensions.

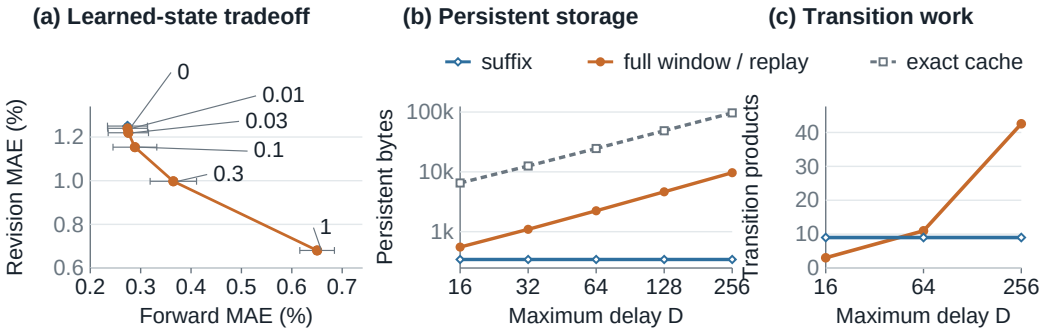

Figure 2: **Representation, storage, and update work.** *(a)* Forward–revision tradeoff for same-size recurrent bottlenecks under six correction-loss weights (mean and Student-t 95% intervals over six seeds); *(b)* full-window versus cut–suffix storage (dashed: exact cache); *(c)* transition products.

6 Computational Study

Realization. At $(K, \epsilon, D) = (16, 2^{-8}, 256)$, the packed cut–suffix state uses 343 bytes versus 9,680 for a same-accuracy full-window grid under identical accounting (28.2× less; Figure 2b). At $K = 8$, materialization uses nine transition products versus 42.56 for full replay (4.73×; Figure 2c). Worst error over 159,825 prefixes is 0.005972ϵ; wall time remains descriptive because the Python bit loop is not optimized.

Learned state. We construct 4,500 pairs of binary-HMM histories with the same terminal belief but different late responses, and keep both members of every pair in the same train, validation, or test split. Same-size one-unit GRUs sweep the correction-loss weight $\lambda \in \{0, .01, .03, .1, .3, 1\}$; identical fresh probes are fit on frozen latents. Every positive weight lowers revision error in all six seeds. At $\lambda = .1$, forward MAE changes from 0.274% to 0.289% while revision MAE falls by 7.7%; at $\lambda = 1$, it falls from 1.25% to 0.68% (45.6%; Figure 2a). Normalized separation between paired latents rises from 0.0316 to 0.0902 ($2.86\times$) across the same sweep. This is a targeted tradeoff diagnostic, not a neural law.

Numerical checks of the constructions. Among 16,128 positive models no sampled response block is singular (84.2% have margin above 10^{-4}). A separate direct-recomputation scan matches all 27,513 checked prefixes with maximum error 0.002764ϵ; enumerations recover the proof ranks and probes. These computations check constructions, not genericity or closed-loop control.

Takeaway and scope. Prediction merges histories with the same future law; revision must separate those that respond differently to evidence inserted into their past. The phase law makes this distinction quantitative and suggests a compact forward state paired with correction memory set by L_ϵ. The theorem assumes a known finite-state model, locally rich bounded evidence, and a response margin; uniform K-growth is proved for the reset family when $\log(1/\epsilon) \geq C_0 \log K$. It is a worst-case correctability contract, not a neural, control-gain, or universal-runtime theorem; average-case memory remains open.

REFERENCES

Yaakov Bar-Shalom. Update with out-of-sequence measurements in tracking: Exact solution. *IEEE Transactions on Aerospace and Electronic Systems*, 38(3):769–777, 2002. doi: 10.1109/TAES.2002.1039398.

Karl Berntorp, Karl-Erik Årzén, and Anders Robertsson. Storage efficient particle filters with multiple out-of-sequence measurements. In *Proceedings of the 15th International Conference on Information Fusion*, pp. 471–478. IEEE, 2012. URL https://lup.lub.lu.se/record/2837293.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Yann Bouteiller, Simon Ramstedt, Giovanni Beltrame, Christopher J. Pal, and Jonathan Binas. Reinforcement learning with random delays. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=QFYnKlBJYR.

Xavier Boyen and Daphne Koller. Tractable inference for complex stochastic processes. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pp. 33–42. Morgan Kaufmann, 1998. URL https://ai.stanford.edu/~xb/uai98/.

Anthony Carbery and James Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters*, 8(3):233–248, 2001. doi: 10.4310/MRL.2001.v8.n3.a1. URL https://intlpress.com/site/pub/files/_fulltext/journals/mrl/2001/0008/0003/MRL-2001-0008-0003-a001.pdf.

Mireille Chaleyat-Maurel and Dominique Michel. Des résultats de non existence de filtre de dimension finie. *Stochastics*, 13(1–2):83–102, 1984. doi: 10.1080/17442508408833312.

Charalambos D. Charalambous and Photios A. Stavrou. On the relation of nonanticipative rate distortion function and filtering theory. In *2013 European Control Conference*, pp. 1627–1632, 2013. URL https://arxiv.org/abs/1210.2019.

Erik D. Demaine, John Iacono, and Stefan Langerman. Retroactive data structures. *ACM Transactions on Algorithms*, 3(2):13:1–13:21, 2007. doi: 10.1145/1240233.1240236.

Randal Douc, Eric Moulines, and Ya'acov Ritov. Forgetting of the initial condition for the filter in general state-space hidden markov chain: A coupling approach. *Electronic Journal of Probability*, 14:27–49, 2009. doi: 10.1214/EJP.v14-593.

Yonathan Efroni, Chi Jin, Akshay Krishnamurthy, and Sobhan Miryoosefi. Provable reinforcement learning with a short-term memory. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5832–5850, 2022. URL https://proceedings.mlr.press/v162/efroni22a.html.

Christopher J. Ellison, John R. Mahoney, and James P. Crutchfield. Prediction, retrodiction, and the amount of information stored in the present. *Journal of Statistical Physics*, 136:1005–1034, 2009. doi: 10.1007/s10955-009-9808-z.

Jamie S. Evans and Vikram Krishnamurthy. Hidden Markov model state estimation with randomly delayed observations. *IEEE Transactions on Signal Processing*, 47(8):2157–2166, 1999. doi: 10.1109/78.774757.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Learning in observable POMDPs, without computationally intractable oracles. In *Advances in Neural Information Processing Systems 35*, 2022. doi: 10.52202/068431-0107. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/099607cd970f4e1ac2fdd30624dffff8-Abstract-Conference.html.

Yanjun Han, Tianze Jiang, and Yihong Wu. Prediction from compression for models with infinite memory, with applications to hidden Markov and renewal processes. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 2270–2307, 2024. URL https://proceedings.mlr.press/v247/han24a.html.

Masahito Hayashi and Vincent Y. F. Tan. Minimum rates of approximate sufficient statistics. *IEEE Transactions on Information Theory*, 64(2):875–888, 2018. doi: 10.1109/TIT.2017.2775612.

Jérémie Houssineau, Ajay Jasra, and Sumeetpal S. Singh. On large lag smoothing for hidden Markov models. *SIAM Journal on Numerical Analysis*, 57(6):2812–2828, 2019. doi: 10.1137/18M1198004.

Qingqing Huang, Rong Ge, Sham M. Kakade, and Munther A. Dahleh. Minimal realization problems for hidden Markov models. *IEEE Transactions on Signal Processing*, 64(7):1896–1904, 2016. doi: 10.1109/TSP.2015.2510969.

Ali Devran Kara and Serdar Yüksel. Near optimality of finite memory feedback policies in partially observed Markov decision processes. *Journal of Machine Learning Research*, 23(11):1–46, 2022. URL https://jmlr.org/papers/v23/20-1152.html.

Armin Karamzade, Kyungmin Kim, JB Lanier, Davide Corsi, and Roy Fox. Model-based reinforcement learning under random observation delays. In *Proceedings of the 8th Annual Learning for Dynamics and Control Conference*, volume 331 of *Proceedings of Machine Learning Research*, pp. 98–116, 2026. URL https://proceedings.mlr.press/v331/karamzade26a.html.

Joona Karjalainen, Anthony Lee, Sumeetpal S. Singh, and Matti Vihola. On the forgetting of particle filters, 2023. URL https://arxiv.org/abs/2309.08517. Version 3, 2026.

Mintae Kim. Finite memory belief approximation for optimal control in partially observable Markov decision processes. *arXiv preprint arXiv:2601.03132*, 2026. URL https://arxiv.org/abs/2601.03132.

Soung Hie Kim and Byung Ho Jeong. A partially observable Markov decision process with lagged information. *Journal of the Operational Research Society*, 38(5):439–446, 1987. doi: 10.1057/jors.1987.72.

François Le Gland and Nadia Oudjane. Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. *The Annals of Applied Probability*, 14(1):144–187, 2004. doi: 10.1214/aoap/1075828050.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems 14*, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/hash/1e4d36177d71bbb3558e43af9577d70e-Abstract.html.

Mahendra Mallick, Thiagalingam Kirubarajan, and Sanjeev Arulampalam. Out-of-sequence measurement processing for tracking ground target using particle filters. In *Proceedings of the 2002 IEEE Aerospace Conference*, volume 4, pp. 4–1809–4–1818, 2002. doi: 10.1109/AERO.2002.1036893.

Sarah Marzen and James P. Crutchfield. Circumventing the curse of dimensionality in prediction: Causal rate-distortion for infinite-order markov processes, 2014. URL https://arxiv.org/abs/1412.2859.

Anthony T. Nixon. The Myhill–Nerode theorem for bounded interaction: Canonical abstractions via agent-bounded indistinguishability, 2026. URL https://arxiv.org/abs/2603.21399.

Boris N. Oreshkin, Xuan Liu, and Mark J. Coates. Efficient delay-tolerant particle filtering. *IEEE Transactions on Signal Processing*, 59(7):3369–3381, 2011. doi: 10.1109/TSP.2011.2140110.

Umut Orguner and Fredrik Gustafsson. Storage efficient particle filters for the out of sequence measurement problem. In *Proceedings of the 11th International Conference on Information Fusion*, pp. 1–8, 2008. doi: 10.1109/ICIF.2008.4632237. URL https://doi.org/10.1109/ICIF.2008.4632237.

Houman Owhadi, Clint Scovel, and Tim Sullivan. On the brittleness of bayesian inference. *SIAM Review*, 57(4):566–582, 2015. doi: 10.1137/130938633.

Pascal Poupart and Craig Boutilier. Value-directed compression of POMDPs. In *Advances in Neural Information Processing Systems 15*, 2002. URL https://proceedings.neurips.cc/paper_files/paper/2002/hash/14ea0d5b0cf49525d1866cb1e95ada5d-Abstract.html.

Nicholas Roy and Geoffrey J. Gordon. Exponential family PCA for belief compression in POMDPs. In *Advances in Neural Information Processing Systems 15*, 2002. URL https://proceedings.neurips.cc/paper/2002/hash/a11f9e533f28593768ebf87075ab34f2-Abstract.html.

Cosma Rohilla Shalizi and James P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104:817–879, 2001. doi: 10.1023/A:1010388907793.

Vatsal Sharan, Sham M. Kakade, Percy Liang, and Gregory Valiant. Prediction with a short memory. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1074–1087, 2018. doi: 10.1145/3188745.3188954.

L. Shue, Brian D. O. Anderson, and Subhrakanti Dey. Exponential stability of filters and smoothers for hidden Markov models. *IEEE Transactions on Signal Processing*, 46(8):2180–2194, 1998. doi: 10.1109/78.705429.

Louis Shue and Subhrakanti Dey. Complexity reduction in fixed-lag smoothing for hidden Markov models. *IEEE Transactions on Signal Processing*, 50(5):1124–1132, 2002. doi: 10.1109/78.995068.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(12):1–83, 2022. URL https://jmlr.org/papers/v23/20-1165.html.

Sekhar Tatikonda and Sanjoy K. Mitter. Control under communication constraints. *IEEE Transactions on Automatic Control*, 49(7):1056–1068, 2004. doi: 10.1109/TAC.2004.831187.

Ramon van Handel. Discrete time nonlinear filters with informative observations are stable. *Electronic Communications in Probability*, 13:562–575, 2008. doi: 10.1214/ECP.v13-1423.

Ramon van Handel. Observability and nonlinear filtering. *Probability Theory and Related Fields*, 145(1–2):35–74, 2009. doi: 10.1007/s00440-008-0161-y.

Shuo Zhang and Yaakov Bar-Shalom. Out-of-sequence measurement processing for particle filter: Exact bayesian solution. *IEEE Transactions on Aerospace and Electronic Systems*, 48(4):2818–2831, 2012. doi: 10.1109/TAES.2012.6324663.

AI USE STATEMENT

We used an AI assistant solely for limited English grammar and wording edits on parts of the manuscript. The AI assistant was not used for experiment design, data collection, model training/evaluation, or coding; all technical content and results were produced and verified by the authors.

REPRODUCIBILITY STATEMENT

Sections 3–5 state the assumptions for every main result, and the appendix contains complete proofs. The anonymous supplementary package provides the study designs, configurations, implementation, tests, prefix-level and aggregate results, and deterministic figure-generation code.

APPENDIX OVERVIEW

Sections A–B formalize exact late-evidence equivalence and the rare-evidence counterexample. Sections C–F develop the response coordinates used for the lower bounds, first abstractly and then in the binary and three-state models. Sections G–H give the quantized construction and its online extension. Section I describes the numerical studies, while Sections J–K provide the growing-state and revealed-input arguments.

A Exact Equivalence Under Late Evidence

For reachable path posteriors π, π', write $\pi \equiv_R \pi'$ when every word legal for both histories gives the same terminal marginal. The empty word implies equality of current marginals. If $\pi \equiv_R \pi'$ and event e is legal for both, every common legal continuation v from the updated pair is the concatenated word ev from the original pair. Hence $U_e\pi \equiv_R U_e\pi'$, so $\equiv_R$ is a right congruence on the external-scheduler interface. The quotient state

$$C^\star(\pi) = [\pi]_{\equiv_R}, \qquad \delta_e[\pi] = [U_e\pi], \qquad d[\pi] = b(\pi)$$

is well defined and exact. Conversely, if a deterministic exact state maps π and π' to one value, deterministic updates and decoding agree after every common word, so $\pi \equiv_R \pi'$. Every exact state therefore refines the quotient. Under an internal-scheduler interface, the equivalence is augmented by equality of the legal continuation languages; the corresponding pending-mask cost is accounted for separately in Theorem 4(a).

For a positive K-state Markov path posterior with known edge potentials, $D+1$ node marginals, each with $K-1$ independent entries, are an exact state. Indeed, the path law is a minimal node-potential exponential family: no nonzero linear combination of independent node indicators is constant on all positive paths. Its log partition is strictly convex, hence its gradient, the node marginals, is one-to-one. These marginals recover the local node potentials; multiplying one local potential by an incoming likelihood and running the known chain recursion gives the updated state. This construction has $(K-1)(D+1)$ real coordinates.

B Rare Evidence and Infinite Packing

The normalizer in Section 4 is finite because $\lambda > c/2$ and is common to both states by the substitution $k \mapsto -k$. Every channel outcome is positive. Conditional on the on-time history $h_n = \{Y = 2n\}$, the posterior log odds of $X = 0$ are $2nc$. For distinct n, m, append the same admissible late outcome $Y' = -(n+m)$. The two posterior log odds are $(n-m)c$ and $(m-n)c$. If $\sigma(t) = (1+e^{-t})^{-1}$, then

$$\Pr(Z = 0 \mid t) = \beta + (1-2\beta)\sigma(t).$$

Their terminal separation is

$$(1-2\beta)\,|\sigma((n-m)c) - \sigma((m-n)c)| = (1-2\beta)\tanh(|n-m|c/2) \geq \delta.$$

Thus $\{h_n : n \geq 1\}$ is an infinite strict packing for every radius below δ. Proposition 1 rules out any finite code when $2\epsilon < \delta$. The separating outcome is rare but legal:

$$\Pr(Y' = -(n+m) \mid h_n) \leq A^{-1}e^{-(\lambda - c/2)(n+m)} \longrightarrow 0.$$

C From Local Observability to a Uniform Packing

We prove the chart claims used by Theorem 3. Lemma 1 gives legal fields $v_1, \ldots, v_m$, $m \leq q$, whose local response derivatives form the full-column-rank stack

$$A_{\mathcal{V}}(r) = \begin{bmatrix} A_{v_1}(r)^\top & \cdots & A_{v_m}(r)^\top \end{bmatrix}^\top.$$

On any compact set of base points, a finite union of such probe families has smallest singular value bounded below. We use that union below; increasing m changes constants but not the rate.

Propagation of local response directions. For a fixed base u, a displacement evolves through

$$T_u(\delta) = F(u+\delta) - F(u).$$

Because F is injective, every T_u is injective. Positivity and bounded fields confine both baseline and finitely probed trajectories to a compact set $\mathcal{R}$. Let

$$\mu = \min\{1, \inf_{r \in \mathcal{R}} \sigma_{\min}(DF(r))\} > 0, \qquad a = \min\{1, \inf_{r \in \mathcal{R}} \sigma_{\min}(A_{\mathcal{V}}(r))\} > 0.$$

At probe age e, vertically stack the m derivatives of terminal-logit displacement. Each block equals $M_{e,j}A_{v_j}(r)$, where $M_{e,j}$ is a product of at most e prediction Jacobians along the corresponding probe trajectory. Hence, for every unit z,

$$\|S_e z\|_2^2 = \sum_{j=1}^{m} \|M_{e,j}A_{v_j}(r)z\|_2^2 \geq \mu^{2e}\|A_{\mathcal{V}}(r)z\|_2^2,$$

and $\sigma_{\min}(S_e) \geq a\mu^e$. At the stationary zero-field trajectory, take a thin SVD $S_e = U_e\Sigma_e V_e^\top$ and set $R_e = U_e^\top$. Then $\|R_e\|_2 = 1$ and the square matrix $R_e S_e$ retains every singular value.

An injective coordinate map. Use the baseline filtering logits $r_0, \ldots, r_L$ as input coordinates. Their local fields satisfy $\theta_0 = r_0 - r_\star$ and $\theta_i = r_i - F(r_{i-1})$ for $i \geq 1$. This coordinate change is triangular with identity diagonal and determinant one. Order inputs as $r_L, r_{L-1}, \ldots, r_0$. The first chart output is the baseline terminal logit. For each increasing age e, append R_e applied to the stacked terminal-logit displacements from the m legal single-message words. An age-e displacement depends on $r_{L-e}, \ldots, r_{L-1}$ but not on r_L; the square chart is therefore reverse block triangular.

On the common compact set, derivatives and second derivatives of an e-step composition are at most e^{Ce}. Choose $b > C + \log(1/\mu)$ and restrict every baseline block to a convex cube of radius

$$\delta_L = ce^{-bL}/(L+1)^2.$$

After reducing c, the induced local fields remain inside the legal on-time neighborhood. A chain-rule telescope bounds every conditional-derivative perturbation on this cube by $C_0(L+1)^2 e^{CL}\delta_L = C_0 c e^{-(b-C)L}$. Taking $c \leq a/(2C_0)$ makes this at most $a\mu^L/2 \leq a\mu^e/2$ for every age $e \leq L$. For one conditional diagonal map G_e and two values x, y of its active block,

$$\begin{aligned} G_e(y) - G_e(x) &= R_e S_e(y - x) \\ &\quad + \int_0^1 [DG_e(x + t(y-x)) - R_e S_e](y - x)\, dt, \end{aligned}$$

so $\|G_e(y) - G_e(x)\|_2 \geq a\mu^e\|y - x\|_2/2$. Each conditional block is thus co-Lipschitz and injective. Starting with the baseline output recovers r_L. Conditional on that block, equality of the age-one outputs and injectivity of its diagonal map recover r_{L-1}. Inductively, after $r_L, \ldots, r_{L-e+1}$ have been recovered, equality of the age-e output and conditional injectivity recover r_{L-e}. Backward substitution therefore recovers every input block and proves global, rather than merely local, injectivity on the product cube. The baseline diagonal block is the identity, whose smallest singular value is at least $a/2$ by the definition of a; the response blocks obey the bounds above. Thus its Jacobian obeys

$$|\det D\Phi_L| \geq \prod_{e=0}^{L} (a\mu^e/2)^q \geq e^{-C_1 L^2}.$$

Packing and metric transfer. Let $n = L + 1$ and $x = \log(1/\epsilon)$. Change of variables on the injective qn-dimensional cube gives response-image volume at least

$$e^{-C_1 L^2}\,[2ce^{-bL}/n^2]^{qn}.$$

Write $z_w(\pi)$ for the terminal reduced logit after legal word w, with $w = \varnothing$ for the baseline. If two full chart outputs are separated by s in Euclidean norm, one of their n blocks is separated by at least $s/\sqrt{n}$. A baseline-block separation already compares the legal empty continuation. Otherwise $\|R_e\|_2 = 1$ shows that one of the m displacement blocks obeys

$$\|[z_{w_j}(\pi) - z_\varnothing(\pi)] - [z_{w_j}(\pi') - z_\varnothing(\pi')]\|_2 \geq s/\sqrt{mn}.$$

The triangle inequality then makes at least one of the two legal raw responses w_j and $\varnothing$ differ in reduced logits by $s/(2\sqrt{mn})$.

It remains to transfer raw-logit separation to probability separation without assuming that a varying Jacobian cannot cancel along a segment. The reference-softmax Jacobian on the first q coordinates is $H(p) = \mathrm{diag}(p_{1:q}) - p_{1:q}p_{1:q}^\top$. For every u,

$$u^\top H(p)u \geq p_K \sum_{i=1}^{q} p_i u_i^2.$$

All terminal logits and their connecting segments lie in a fixed compact set, so every probability coordinate is at least some $\underline{p} > 0$ there. Integrating the displayed positive-definite bound along a segment and applying Cauchy–Schwarz gives

$$\|\sigma(r')_{1:q} - \sigma(r)_{1:q}\|_2 \geq \underline{p}^{\,2}\|r' - r\|_2.$$

Total variation is at least half the Euclidean distance between the first q probability coordinates. Thus the preceding legal raw-response separation converts to response-metric separation by a fixed model-dependent factor. We may pack the chart image at Euclidean radius $C\epsilon\sqrt{mn}$ to obtain a strict 2ϵ packing in d_R. The $\sqrt{mn}$ rescaling contributes only $O(qn \log n)$ to the volume estimate, so a maximal strict response packing satisfies

$$\log \mathcal{P} \geq qnx - C_2 n^2 - O(qn \log n).$$

Choose $n = \min\{D + 1, \max(1, \lfloor \gamma x \rfloor)\}$ for sufficiently small fixed $\gamma > 0$. After reducing ϵ_0, the n^2, $qn \log n$, and linear terms together consume at most a fixed fraction of qnx, uniformly for $1 \leq n \leq \gamma x$. Proposition 1 therefore yields $\Omega(q \min\{D + 1, x\}x)$ bits. Theorem 2 applies because the bounded delayed set lies in a fixed field box, proving the matching all-regime upper bound.

With $L = D$, the same chart is continuously injective on an open $q(D + 1)$-dimensional patch. Invariance of domain gives the exact lower bound. The positive node-potential family is minimal, so its $q(D + 1)$ independent node marginals recover the local fields and attain the exact dimension.

All inequalities used above are strict on the base chart and continuous in the model and field coordinates. Hence, for each finite D, a smaller open product cube retains injectivity, a positive Jacobian margin, and response separation under all sufficiently small parameter perturbations. The cube has positive Lebesgue measure in the declared on-time field support. This proves Corollary 2 and makes explicit that the packing is not concentrated on an isolated observation history.

D Observability Beyond the Reset Witnesses

We prove Proposition 2. Fix K, put $q = K - 1$, and let $\sigma : \mathbb{R}^q \to \operatorname{int} \Delta_K$ and $\lambda : \operatorname{int} \Delta_K \to \mathbb{R}^q$ be reference-class softmax and reduced logits. Because P is full rank, its restriction to the simplex hyperplane is injective. Hence $F_P = \lambda \circ P^\top \circ \sigma$ is injective and $DF_P(r)$ is nonsingular everywhere. Positivity maps the simplex into a compact subset of its interior and gives a Birkhoff coefficient $\tau(P) < 1$ in log oscillation.

Local response coordinates. Retain $n = L + 1$ coordinates. For the on-time field blocks $\theta = (\theta_0, \ldots, \theta_L)$, form a qn-dimensional chart from the terminal reduced logits before a late message and the terminal-logit displacements after inserting v separately at each of the L older coordinates. Order input blocks newest to oldest and responses by increasing probe age. At zero field the baseline is $r_\star$. Write

$$A = DH_{P,v}(r_\star) = DF_P(r_\star + v) - DF_P(r_\star), \quad x_0 = r_\star + v, \quad x_{t+1} = F_P(x_t).$$

The chart is block triangular. Its baseline diagonal block is the identity, and the age-e diagonal block is

$$G_e = DF_P(x_{e-1}) \cdots DF_P(x_1)A. \tag{2}$$

Thus its Jacobian at zero is nonsingular under (OC).

We need more than local rank. The inverse-function theorem gives a fixed neighborhood $\mathcal{N}$ of $r_\star$ on which $H_{P,v}(r) = F_P(r + v) - F_P(r)$ is one-to-one. For fixed u, every subsequent displacement transport

$$T_u(\delta) = F_P(u + \delta) - F_P(u)$$

is one-to-one because F_P is. Conditional on newer baseline blocks, the age-e diagonal response is a composition of $H_{P,v}$ and such transports. Equality of two complete charts therefore identifies the oldest differing block; backward block substitution identifies every block. The chart is globally one-to-one whenever all baselines entering $H_{P,v}$ lie in $\mathcal{N}$.

A uniform neighborhood and packing bound. Take each field block in a cube of radius c_0/n^2. In any fixed norm on reduced logits, Birkhoff contraction and local norm equivalence give

$$\|r_i - r_\star\| \le \tau\|r_{i-1} - r_\star\| + Cc_0/n^2, \qquad \tau < 1.$$

For sufficiently small model-dependent c_0, every baseline lies in $\mathcal{N}$. Positivity also confines every probe orbit and its perturbations to a second fixed compact set. Continuity and (OC) then give fixed $c_A, \eta > 0$ such that $|\det DH_{P,v}| \ge c_A$ and $|\det DF_P| \ge \eta$ on the required sets. The change-of-variables formula and the block determinant yield

$$\mathrm{Vol}(\Phi_L(\Theta_L)) \ge (2c_0/n^2)^{qn} c_A^L \eta^{L(L-1)/2}. \tag{3}$$

Reduced-logit and terminal-probability Jacobians are uniformly bounded above and below there, changing (3) by only an exponential factor in n.

Let $x = \log(1/\epsilon)$ and $\beta = -\log\min\{\eta, 1\}$. A maximal strict 2ϵ packing in qn dimensions obeys

$$\log \mathcal{P} \ge qnx - \frac{\beta}{2}L(L-1) - 2qn\log n - O(n).$$

Choose $n = \min\{D{+}1, \max(2, \lfloor \gamma x \rfloor)\}$ with sufficiently small fixed $\gamma > 0$; if $\beta = 0$, any sufficiently small γ works. Shrinking ϵ_0 ensures the cube is legal. Uniformly for $2 \le n \le \gamma x$, the determinant and cube costs are a fixed fraction of qnx. For the finitely many cases where only one block is retained, the baseline-output chart supplies the same $\Omega(qx)$ order with model-dependent constants. Proposition 1 therefore gives $\Omega(q\min\{D{+}1, x\}x)$ bits, and Theorem 2 gives the matching all-regime upper bound.

With $L = D$, the same full-rank chart is continuously injective on an open $q(D+1)$-dimensional patch. Every continuous exact encoder must be injective there, so invariance of domain gives the exact-dimension lower bound. For the upper bound, the positive path posterior is a node-potential exponential family over the fixed Markov base measure. Removing one reference-state indicator per timestamp leaves $q(D+1)$ minimal statistics: full path support rules out a nontrivial linear combination constant on all paths. Strict convexity of the log partition makes their node marginals one-to-one with the local fields, so they reconstruct the posterior and attain dimension $q(D+1)$.

A positive-measure class for every fixed K. For the reset family with coefficient α and a categorical outcome with boost $t = e^a$, direct differentiation makes A triangular. Let

$$d_1 = \frac{Kt}{t(1+\alpha(K-1)) + (1-\alpha)(K-1)}, \qquad d_o = \frac{K}{(1-\alpha)t + K - 1 + \alpha}.$$

The eigenvalues of A are

$$\alpha\left[\frac{d_1(K-1) + td_o}{t+K-1} - 1\right] \quad \text{and} \quad \alpha(d_o - 1) \quad (K-2 \text{ times}).$$

At $\alpha = 1/2, t = 2$, these reduce to

$$\frac{2K^2 - 5K - 1}{2(2K+1)(3K+1)}, \qquad -\frac{1}{2(2K+1)}.$$

The repeated eigenvalue is nonzero. The first numerator has discriminant 33 and no integer root, so it is nonzero for every integer $K \ge 2$. Condition (OC) therefore holds at every declared reset witness.

We now strengthen this witness argument to an almost-everywhere statement. Use the first q entries of each row as coordinates for

$$\mathcal{P}_K^+ = \{P : P_{ij} > 0,\ P\mathbf{1} = \mathbf{1}\},$$

so $\mathcal{P}_K^+$ is a connected open convex subset of $\mathbb{R}^{Kq}$. A positive binary late channel is parameterized by $\ell = (\ell_1, \dots, \ell_K) \in (0,1)^K$, where $\ell_s = \Pr(W = 1 \mid X = s)$. Its distinguished outcome has reduced field

$$v(\ell)_i = \log(\ell_i/\ell_K), \qquad 1 \le i \le q.$$

Conversely, every finite $v \in \mathbb{R}^q$ is legal: put $\widetilde{v} = (v_1, \dots, v_q, 0)$ and choose $0 < c < \exp(-\max_s \widetilde{v}_s)$. Then $\ell_s = c\exp(\widetilde{v}_s) \in (0,1)$ and the binary channel with rows $(\ell_s, 1-\ell_s)$ realizes exactly v.

Let $r_\star(P) = \lambda(\pi(P))$ and define

$$\Psi_K(P, \ell) = \det[DF_P(r_\star(P) + v(\ell)) - DF_P(r_\star(P))] .$$

This function is real analytic on $\mathcal{U}_K = \mathcal{P}_K^+ \times (0,1)^K$. For the only non-obvious term, replace the last row of $I - P^\top$ by $\mathbf{1}^\top$ to form $M(P)$. Because $\mathbf{1}^\top(I - P^\top) = 0$, the system $M(P)\pi = e_K$ is equivalent to stationarity plus normalization. Positivity gives a unique stationary probability, hence $M(P)$ is invertible. Cramer's rule makes each coordinate of $\pi(P)$ a ratio of polynomials with a nowhere-zero denominator. The remaining operations defining Ψ_K preserve real analyticity.

For completeness, a nonzero real-analytic f on a connected open subset of $\mathbb{R}^m$ has a null zero set with empty interior. At each zero x, choose $D^\alpha f(x) \neq 0$ of minimal order, choose j with $\alpha_j > 0$, and put $\beta = \alpha - e_j$. Minimality puts x in $\{D^\beta f = 0, \partial_j D^\beta f \neq 0\}$, locally a C^1 hypersurface. A countable union over multi-indices and coordinate charts is null; the analytic identity theorem rules out interior.

At $P_K = \frac{1}{2}I + \frac{1}{2K}\mathbf{1}\mathbf{1}^\top$ and $\ell = (1/2, 1/4, \ldots, 1/4)$, the field is $(\log 2, 0, \ldots, 0)$ and the eigenvalue calculation above gives $\Psi_K(P_K, \ell) \neq 0$. Also $\det(P_K) = 2^{-(K-1)}$. Hence the analytic function

$$\mathcal{R}_K(P, \ell) = \det(P)\Psi_K(P, \ell)$$

is not identically zero on the connected domain $\mathcal{U}_K$. Its nonzero set is therefore open, dense, and of full Lebesgue measure in the stated coordinates. On exactly that set, P is full rank and (OC) holds, so the preceding chart proof applies. This proves Proposition 2. Genericity here is relative to this declared positive transition–binary-channel family for a fixed K; it neither covers arbitrary HMM parameterizations nor supplies constants uniform in growing K.

Uniform constants away from singular parameters. We prove Proposition 3. In free row-stochastic coordinates,

$$DF_P(\lambda(p)) = D\lambda(P^\top p)P^\top D\sigma(\lambda(p)).$$

Its entries are rational functions of (P, p). Cramer's rule makes the stationary probability $\pi(P)$ rational, and the probed probability is

$$p_s^\ell = \frac{\pi_s(P)\ell_s}{\sum_j \pi_j(P)\ell_j}.$$

After clearing the finitely many positive denominators, therefore,

$$\mathcal{R}_K(P, \ell) = \frac{N_K(P, \ell)}{S_K(P, \ell)},$$

where N_K, S_K are polynomials, $S_K > 0$ on $\mathcal{U}_K$, and N_K is not the zero polynomial because the reset witness has $\mathcal{R}_K \neq 0$. It is not constant either: a positive transition with identical rows has $\det P = 0$ and hence $\mathcal{R}_K = N_K = 0$. Put $d_K = \deg N_K \geq 1$. For a full-dimensional compact convex $\mathcal{C} \Subset \mathcal{U}_K$, its normalized Lebesgue measure is log-concave. The Carbery–Wright small-ball inequality (Carbery & Wright, 2001) and $S_{\max} = \max_{\mathcal{C}} S_K$ give

$$\mu_{\mathcal{C}}(|\mathcal{R}_K| \leq \gamma) \leq c_{\mathrm{CW}} d_K \left(\frac{S_{\max}\gamma}{\|N_K\|_{L^2(\mu_{\mathcal{C}})}} \right)^{1/d_K}.$$

Indeed, $|\mathcal{R}_K| \leq \gamma$ implies $|N_K| \leq S_{\max}\gamma$; moreover $\|N_K\|_{L^2(\mu_{\mathcal{C}})} > 0$ because a nonzero polynomial cannot vanish on the interior of $\mathcal{C}$. Truncating the right side at one proves the stated small-margin bound.

It remains to justify the uniform matching-law constants. On the compact set $\mathcal{C}_\gamma = \mathcal{C} \cap \{|\mathcal{R}_K| \geq \gamma\}$, all matrix norms and denominators are uniformly bounded. Since $|\det P \det A| \geq \gamma$, upper bounds on each determinant give positive lower bounds on the other; for a $q \times q$ matrix,

$$\sigma_{\min}(M) \geq \frac{|\det M|}{\|M\|_2^{q-1}}.$$

Thus P and the response block are uniformly nonsingular. Compact positivity also gives a common Birkhoff contraction, bounded probe orbit, derivative bounds, and a quantitative inverse-function

neighborhood. Together with the assumed common on-time field neighborhood, these yield uniform c, C, C_0, ϵ_0 in the chart and quantization arguments on $\mathcal{C}_\gamma$.

Finally, for the reset witness with $K \geq 3$, the response matrix has repeated eigenvalue $-1/[2(2K+1)]$. If x is its corresponding eigenvector, then $\sigma_{\min}(A_K) \leq \|A_K x\|_2/\|x\|_2 = 1/[2(2K+1)]$. Hence this family has no K-independent Euclidean response margin in the declared reduced-logit coordinates. This is not a coordinate-invariant impossibility theorem and does not weaken the fixed-K result; it makes explicit why we claim no growing-K uniformity.

E The Binary Witness

We give details for the binary determinant identity in Section 5. Write $m = \mathbb{E}[S_L]$, so the terminal probability is $(1+m)/2$. At zero field, $\mathbb{E}[S_i S_j] = \rho^{|i-j|}$. Under the single probe at age e,

$$\mathbb{E}[S_L] = \tanh(a)\rho^e, \quad \mathbb{E}[S_{L-d}] = \tanh(a)\rho^{|d-e|}, \quad \mathbb{E}[S_L S_{L-d}] = \rho^d.$$

Differentiating an exponential tilt yields a covariance. If $C_L = 2D\Phi_L(0)$, its baseline row is $(C_L)_{0d} = \rho^d$ and its probed row is

$$(C_L)_{ed} = \rho^d - q\rho^e\rho^{|d-e|}, \qquad q = \tanh^2 a.$$

Factor ρ^d from column d, subtract the baseline row from each probed row, then take consecutive differences in probe age. The resulting matrix is triangular with diagonal magnitude $q(1-\rho^2)$. Consequently

$$|\det D\Phi_L(0)| = 2^{-n}[q(1-\rho^2)]^L \rho^{\sum_{d=0}^{L} d}.$$

It remains to justify a uniform injective patch rather than only a nonzero point determinant. The binary filter log odds satisfy

$$r_i = 2\theta_i + \varphi(r_{i-1}), \qquad \varphi(r) = 2\,\mathrm{arctanh}(\rho\tanh(r/2))\,.$$

Apply logit coordinatewise to all chart outputs; on the shrinking field cube, the output probabilities remain in a fixed compact subset of $(0,1)$, so the logit Jacobian is bounded above and below by constants to the power n. Subtract the baseline logit from the probed logits, a determinant-one row operation.

For a probe at coordinate $j = L - e$, let Δ_k be its log-odds displacement at time k. Then

$$\Delta_j = 2a, \qquad \Delta_{k+1} = \varphi(r_k + \Delta_k) - \varphi(r_k).$$

In reverse parameter order, the response map is triangular. Holding later reverse coordinates fixed, the exact diagonal derivative at zero has magnitude

$$\frac{\rho^e q(1-\rho^2)}{1 - q\rho^{2e}} \geq \rho^e q(1-\rho^2).$$

It has a fixed sign. Since $|\varphi'(r)| \leq \rho$ and the field cube has side $O(n^{-2})$, the accumulated baseline displacement is $O(n^{-1})$ and the relative perturbation of the derivative product is bounded uniformly. Choose c_0 so this perturbation is at most one half. Each triangular coordinate is then strictly monotone conditional on later coordinates; backward substitution gives global injectivity. Multiplying the uniform diagonal bounds and undoing the bounded logit transform yields $|\det D\Phi_L| \geq c_1^n \rho^{L(L+1)/2}$.

For exact dimension, take $L = D$. The chart factors through every exact state and is injective on an open $(D+1)$-dimensional patch. If a continuous exact encoder landed in $\mathbb{R}^d$ with $d < D+1$, composing with the standard inclusion into $\mathbb{R}^{D+1}$ would give a continuous injection from an open subset of $\mathbb{R}^{D+1}$ into a lower-dimensional hyperplane, contradicting invariance of domain. The node-marginal construction in Appendix A attains $D+1$.

For the approximate lower bound, all logs in this paragraph are natural. Let $x = \log(1/\epsilon)$ and choose

$$n = L + 1 = \min\{D+1, \max(1, \lfloor \gamma x \rfloor)\},$$

where $\gamma > 0$ is smaller than a fixed multiple of $1/(-\log\rho)$. Choose ϵ_0 so the cube lies within the legal sensor domain. Change of variables gives

$$\mathrm{Vol}(\Phi_L(\Theta_L)) \geq (2c_0/n^2)^n c_1^n \rho^{L(L+1)/2}.$$

A maximal strict 2ϵ packing covers the image by cubes of side 4ϵ, hence

$$\log \mathcal{P} \geq n \log(1/\epsilon) - \frac{-\log \rho}{2} L(L+1) - 2n \log n - O(n) = \Omega(nx) = \Omega(\min\{D+1, x\}x).$$

For the upper bound, let $L_\epsilon = \lceil \log(2/\epsilon)/(-\log \rho) \rceil$. If $D < L_\epsilon$ retain all $D+1$ fields and quantize them at spacing $O(\epsilon/(D+1)^2)$. The normalized-tilt calculation below gives error at most ϵ and $O((D+1)x)$ bits because $D = O(x)$. In the remaining branch, the binary Doob calculation gives truncation error at most $\rho^{L_\epsilon} \leq \epsilon/2$ under arbitrary retained evidence. The cut boundary field is bounded by $\operatorname{arctanh} \rho$ and each retained local field by $h + a$. Project and round the $L_\epsilon + 1$ fields in that box with $\Delta \leq 2\epsilon/(L_\epsilon+1)^2$. With at most $M \leq L_\epsilon$ retained delayed messages,

$$\|\widehat{\theta} - \theta^{\text{ignore}}\|_\infty \leq (M+1)\Delta/2.$$

The path log-weight ratio has oscillation at most $2(L_\epsilon+1)\|\widehat{\theta} - \theta^{\text{ignore}}\|_\infty$; normalized tilting therefore gives

$$\mathrm{TV}(b^{\text{ignore}}, b^{\text{quant}}) \leq \tanh\left(\frac{(L_\epsilon+1)(M+1)\Delta}{4}\right) \leq \epsilon/2.$$

The full/ignore/quantized triangle inequality proves total error at most ϵ. Counting grid indices gives

$$(L_\epsilon+1)\left\lceil \log_2\left(1 + \left\lceil \frac{2(h+a+\operatorname{arctanh}\rho)}{\Delta} \right\rceil\right)\right\rceil = O(\log^2(1/\epsilon)).$$

Taking the appropriate branch gives $O(\min\{D+1, x\}x)$ bits.

F THE THREE-STATE WITNESS

The on-time sensor domain is

$$\Omega_h = \{z \in \mathbb{R}^3 : \mathbf{1}^\top z = 0, \max_s z_s - \min_s z_s < h\},$$

with two-dimensional Lebesgue measure. It is invariant under state permutations, so $p(z \mid X = s) = e^{z_s}/Z_h$ has a common normalizer and reduced fields contain an open two-dimensional neighborhood.

Because softmax and reduced logits are diffeomorphisms and P is nonsingular on the simplex tangent plane, $F = \lambda \circ P^\top \circ \sigma$ is globally injective and DF is nonsingular everywhere. Connectedness and $DF(0) = I/2$ imply $\det DF > 0$. Direct differentiation at $v = (\log 2, 0)$ gives

$$DF(v) = \begin{pmatrix} 18/35 & -3/70 \\ 0 & 3/7 \end{pmatrix}, \qquad A = DF(v) - DF(0) = \begin{pmatrix} 1/70 & -3/70 \\ 0 & -1/14 \end{pmatrix}.$$

Let the baseline recursion be $r_i = \theta_i + F(r_{i-1})$. Under a probe at j, the initial displacement is v and later displacement transport is $T_u(\delta) = F(u+\delta) - F(u)$. At zero, the age-$e$ diagonal block is

$$H_e = DF(F^{e-1}(v)) \cdots DF(F(v))A.$$

Thus the reverse-coordinate chart is block triangular and

$$|\det DG_L(0)| = |\det A|^L \prod_{t=1}^{L-1} |\det DF(F^t(v))|^{L-t}.$$

For global injectivity on a patch, define $H(r) = F(r+v) - F(r)$. Since $DH(0) = A$ is nonsingular, the inverse-function theorem provides a fixed neighborhood on which H is one-to-one. Every subsequent T_u is one-to-one because F is. Small baseline fields stay in the H neighborhood by the local contraction $DF(0) = I/2$, while probe trajectories stay in a fixed compact set. Backward block substitution gives global chart injectivity. Compactness also gives constants $c_1 > 0$ and $0 < \eta < 1$ such that on $[-c_0/n^2, c_0/n^2]^{2n}$,

$$|\det D\Phi_L^{(3)}| \geq c_1^n \eta^{L(L-1)/2}.$$

The chart with $L = D$ forces exact dimension at least $2(D + 1)$; two node marginals per timestamp attain it. For the approximate packing, put $\beta = -\log \eta$, $x = \log(1/\epsilon)$, and choose $n = L + 1 = \min\{D + 1, \max(1, \lfloor \gamma x \rfloor)\}$ for sufficiently small $\gamma > 0$. The $2n$-dimensional image volume obeys

$$\mathrm{Vol}(\Phi_L^{(3)}(\Theta_L^{(3)})) \geq (2c_0/n^2)^{2n} c_1^n \eta^{L(L-1)/2}.$$

Therefore

$$\log \mathcal{P} \geq 2nx - \frac{\beta}{2} L(L-1) - 4n \log n - O(n) = \Omega(nx) = \Omega(\min\{D + 1, x\}x).$$

For the upper bound, every entry of P lies in $[1/6, 2/3]$, giving common Doob minorization $1/4$ and row diameter $3/4$. The cut logits are bounded by $\log 4$, retained fields by $h + a$, and projection–rounding with $\Delta \leq 2\epsilon/(L_\epsilon + 1)^2$ repeats the binary argument in two coordinates after the mixing horizon. Before it, the full-window branch retains all $D + 1$ coordinates. The fixed-lag code length is

$$2(L_\epsilon + 1) \left\lceil \log_2 \left(1 + \left\lceil \frac{2(\log 4 + h + a)}{\Delta} \right\rceil \right) \right\rceil = O(\log^2(1/\epsilon)),$$

and the full-window code costs $O(2(D + 1)x)$. Together they give $O(2 \min\{D + 1, x\}x)$.

G Finite-State Quantization

Prediction across one positive transition puts every cut probability between m and M, so reduced cut logits are bounded by $\log(M/m)$. With initial bound R_0, set $\Lambda = \max\{R_0, \log(M/m)\} + H + A$. The legal exact truncated trajectory stays in $[-\Lambda, \Lambda]^{K-1}$. Projection onto this box is therefore nonexpansive relative to that exact point. One initialization rounding and one possible post-message rounding give per-coordinate error at most Δ.

If $D < L_\epsilon$, apply the construction to all $D+1$ field blocks with $\Delta = 2\epsilon/(D+1)$. The telescope below gives terminal error at most ϵ. There are $q(D+1)$ grid indices, each using $O(\log(1/\epsilon)+\log(D+1)) = O(\log(1/\epsilon))$ bits in this regime. This is the pre-mixing branch of Theorem 2.

To telescope, change one timestamped reduced field at a time while holding all other path factors fixed. The two full path posteriors are normalized tilts of the same positive path law, and a sup-norm field perturbation δ gives their log-density ratio oscillation at most 2δ. Their joint TV is therefore at most $\tanh(\delta/2) \leq \delta/2$; passing to the terminal marginal cannot increase TV. With $\Delta = \epsilon/(L_\epsilon + 1)$,

$$\mathrm{TV}(b^{\mathrm{ignore}}, b^{\mathrm{quant}}) \leq \frac{1}{2} \sum_{d=0}^{L_\epsilon} \Delta \leq \epsilon/2.$$

Combining this with the Doob truncation bound proves the post-mixing branch of Theorem 2 for every subset, permutation, and prefix. The two branches together yield $O(q \min\{D+1, \log(1/\epsilon)\} \log(1/\epsilon))$ bits.

For contraction-matched grids, assign $\Delta_d = \min\{2\Lambda, \epsilon/((L_\epsilon+1)\kappa^d)\}$. Telescope with the remaining d-edge contraction to bound each coordinate's terminal contribution by $\kappa^d \Delta_d/2$. The sum is at most $\epsilon/2$. Summing the uncapped grid lengths gives leading explicit code term

$$\frac{K-1}{2|\log \kappa| \log 2} \log^2(1/\epsilon) + O((K-1)\log(1/\epsilon)\log\log(1/\epsilon)).$$

This coefficient compares two constructive upper codes and is not asserted to be the minimax constant.

H Online Correction Memory

We prove Theorem 4(a). At the current time, the D source timestamps of ages $0, \ldots, D-1$ each have one generated packet that is either received or pending. Because every delay in $\{0, \ldots, D\}$ has positive probability and delays are independent, every pending mask in $\{0, 1\}^D$ is reachable. Packet values are generated by the delayed channel and remain hidden until arrival. The accuracy requirement applies after each individual arrival in a simultaneous bundle, so the same proof handles arbitrary bundle order. For Theorem 4(a), the scheduler exposes the current pending set and the packet being processed; neither is part of the persistent correction code.

Constructive upper bound. Choose $L = \min\{D, \lceil C_\kappa \log(C/\epsilon)\rceil\}$. If $L = D$, the correction state retains and quantizes the whole delay window. Otherwise the common conditional row diameter through L retained edges is at most $\epsilon/3$, and the state consists of a quantized cut belief plus the quantized reduced local field at each retained timestamp. A late packet in the retained window is added to its field block and that block is rounded once; an older packet is ignored only in the second branch. On a time shift, the oldest block is absorbed into the cut update, the remaining grid indices shift without re-rounding, and the new on-time field is rounded and appended.

Compare three pairs of trajectories. First, full processing and exact ignore-old processing agree when $L = D$. Otherwise they differ only in the mixing distribution at the retained cut. Under any common later evidence, the retained conditional channel has row diameter at most $\epsilon/3$, hence

$$\mathrm{TV}(b_t^{\mathrm{full}}, b_t^{\mathrm{old}}) \le \epsilon/3.$$

Second, hold the exact cut boundary fixed and replace retained fields one at a time. With field-grid spacing Δ, initialization plus one post-message round gives coordinate error at most Δ. Holding the other path factors fixed, exact and perturbed path posteriors are normalized local tilts with log-density-ratio oscillation at most 2Δ. Their joint TV is at most $\tanh(\Delta/2) \le \Delta/2$, and terminal marginalization cannot increase it. Thus choosing $\Delta \le 2\epsilon/[3(L+1)]$ yields

$$\mathrm{TV}(b_t^{\mathrm{old}}, b_t^{\mathrm{field}}) \le \epsilon/3.$$

It remains to control repeated cut-boundary quantization. For positive probabilities define Hilbert distance $h(p, p') = \mathrm{osc}(\log(p/p'))$. A local likelihood field ϕ followed by prediction is induced projectively by $P^\top \mathrm{diag}(e^\phi)$. Its cross ratios equal those of $P^\top$, because the diagonal factors cancel. Birkhoff's theorem therefore gives a common $\tau(P) < 1$. Comparing exact and quantized fields before the common positive map and then projection-rounding the boundary at reduced-logit spacing Δ_b gives constants C_1, C_2 such that

$$h_{t+1} \le \tau h_t + C_1\Delta + C_2\Delta_b. \tag{4}$$

Positive prediction puts every cut belief in a compact simplex subset, so its reduced logits lie in a fixed box and projection-rounding is nonexpansive relative to the exact boundary. This recurrence includes the quantization error in the block being dropped; it is not counted only in the retained suffix. Iteration gives

$$h_t \le \tau^t h_0 + \frac{C_1\Delta + C_2\Delta_b}{1-\tau}.$$

Initialize the cut by the same projection–rounding rule, so h_0 is bounded by another constant multiple of Δ_b. Take $\Delta_b \le c_b(1-\tau)\epsilon$ and reduce the constant in Δ so the stationary bound in (4) is at most $4\,\mathrm{arctanh}(\epsilon/3)$. Common suffix likelihoods preserve the oscillation of boundary density ratios, and marginalization cannot increase it. Since $\mathrm{TV}(p, p') \le \tanh(h(p, p')/4)$,

$$\mathrm{TV}(b_t^{\mathrm{field}}, b_t^{\mathrm{quant}}) \le \epsilon/3$$

uniformly over time. The triangle inequality proves current-belief error at most ϵ after every update prefix.

The field bank costs

$$O\left(q(L+1)\log\frac{L+1}{\epsilon}\right) = O(q\min\{D+1, \log(1/\epsilon)\}\log(1/\epsilon))$$

bits, and the boundary costs $O(q\log(1/\epsilon))$. This proves the upper bound.

Lower bound and scheduler metadata. Freeze any online time and choose a legal scheduler realization in which the suffix packets needed by the fixed-time chart remain pending. Accuracy after every arrival prefix then contains the fixed-time response problem as a restriction. The all-regime packing in Theorem 3 gives $\Omega(q\min\{D+1, \log(1/\epsilon)\}\log(1/\epsilon))$ correction bits.

For the internal-scheduler interface, store the pending mask alongside the upper code, giving the additive D-bit upper bound. For the converse, put $x = \log(1/\epsilon)$ and $L' = \max\{1, \lfloor \gamma x \rfloor\}$ for the sufficiently small model-dependent constant γ in the fixed-time packing proof. Suppose $D \ge 2L'$. Fix the newest L' packets pending and vary the older $D - L'$ mask bits. State-independent full-support

delays make all $2^{D-L'}$ masks reachable independently of the latent and sensor outcomes. For any such reachable prefix, positivity of the transition and the bounded local fields place the filtering belief at the cut in one common compact subset of the simplex interior.

The finite-probe argument of Appendix C, applied over this compact cut-belief set and the common open on-time field neighborhood, supplies uniform derivative, probe-margin, and cube-radius constants. Consequently, for every fixed older mask m there is a recent-window response packing $\mathcal{H}_m$ of size at least $\exp(cq(L'+1)\log(1/\epsilon))$, with the same $c > 0$ for every m. Histories within one $\mathcal{H}_m$ cannot share a state by Proposition 1; histories belonging to different masks cannot share one because the interface must recover the exact mask. Summing the packing sizes over masks gives $2^{D-L'}\exp(cq(L'+1)\log(1/\epsilon))$ states. In the complementary regime, the correction lower bound already dominates D for sufficiently small ϵ. Thus the internal state has the additive order in Theorem 4(a); the product argument is used only for that interface.

I Numerical Study and Implementation Details

I.1 Finite-query enumeration and streaming study

The machine-readable specification accompanying the supplement records every parameter, seed, and evaluation rule used below. Binary packing histories are Cartesian products across $D+1$ timestamps of the field set

$$\{-0.12, -0.06, 0, 0.06, 0.12\}.$$

Three-state histories use the four blocks

$$\{(-.05, -.05), (-.05, .05), (.05, -.05), (.05, .05)\}.$$

For each pair of histories, we compute the maximum TV across the baseline and all single delayed outcomes at every strictly past timestamp. A binary variable selects each candidate, and every pair at distance at most 2ϵ produces an incompatibility constraint $z_i + z_j \le 1$. SciPy/HiGHS solves the resulting program with zero requested MIP gap and must return global-optimum status. The reported set is then checked pairwise from the saved distance matrix.

The streaming study crosses three correlations, four maximum delays, three accuracies, and eight fixed seeds. Every setting runs 600 source-time shifts. The full augmented reference, exact fixed-lag filter, and quantized recursive filter share latent paths, on-time fields, delayed values, delays, and bundle orders. Beliefs are recorded before each bundle and after every arrival, giving 28 packing instances, 576 method–seed summaries, 682,722 prefix comparisons, and 21 finite-regime certificates. The largest observed quantized error is 0.068876ϵ.

Correction indices, pending-mask bits, and dense-table storage are accounted for separately. Floating-point workspace is not charged as a compact stored state. The finite-grid optimum is exact for the declared grid and query family; it is a lower bound, not an estimator, for the continuous minimax code.

I.2 Finite-precision constructions

The memory experiment contains 57 model–delay cells: binary $\rho \in \{0.30, 0.60, 0.85\}$ with $D \in \{8, 16, 32, 64, 128, 256\}$; three-state $\alpha \in \{0.35, 0.50, 0.70\}$ with $D \in \{8, 16, 32, 64, 128\}$; and $K \in \{5, 8\}$, $\alpha \in \{0.35, 0.50\}$ with $D \in \{8, 16, 32, 64, 128, 256\}$. Four target accuracies, 24 nested history seeds, five arrival schedules, and five methods produce 136,800 unique cluster rows. Each row summarizes $D+1$ prefixes, totaling 10,389,600 method-prefix scores.

The five schedules are oldest-first cyclic outcomes, newest-first contradictory outcomes, alternating endpoints with maximum-field outcomes, and two independent random permutation/outcome words. The five methods are full replay, timestamp-blind insertion, exact fixed lag, uniform quantization, and age-adaptive quantization. A separate $L = D$ reconstruction is compared with direct full replay.

All 136,800 model–delay–accuracy–history–schedule–method combinations are present, and each contains the expected $D+1$ prefixes. For each of the 80 model–accuracy–quantizer groups, at least two tested delays satisfy $D \ge L_\epsilon$; their stored-bit counts are identical, as predicted by the cut–suffix construction. We also recompute the grid sizes from their integer formulas, verify that every sensor

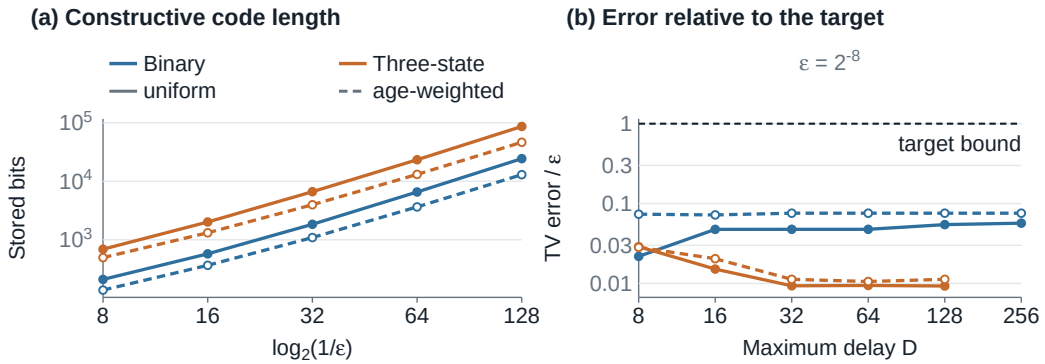

Figure 3: **Finite-precision constructions.** *(a)* Stored bits for the binary and three-state constructions as the target accuracy tightens. Color identifies the model; line style identifies the grid allocation. *(b)* Worst sampled total-variation error relative to the requested tolerance. These calculations illustrate the explicit codes; the minimax rate and uniform error guarantee are analytical.

field lies in the declared support, and compare the $L = D$ recursion directly with full replay. The ignored-prefix, quantization, and total errors are reported separately.

The local response study uses 480 perturbations around the reset witnesses. To examine whether comparable conditioning persists away from these anchors, we additionally cross seven state counts $K \in \{2, 3, 4, 5, 8, 12, 16\}$, four transition geometries (dense independent rows, reversible symmetric weights, directed cyclic structure, and two-block structure), three row-logit scales $\{0.25, 1, 2.5\}$, three delayed-field bounds $\{0.25, 1, 2.5\}$, and 64 deterministic samples per setting, for 16,128 models in total.

For every model we record transition positivity, stationary and channel margins, contraction, transition conditioning, the response-block singular spectrum and log determinant, and a length-four cocycle determinant. One model per setting is also checked by finite differences. The sampled $\sigma_{\min}$ minimum, 1st percentile, 5th percentile, and median are respectively 1.20×10^{-8}, 2.48×10^{-6}, 1.69×10^{-5}, and 1.38×10^{-3}; 99.53% exceed 10^{-6}. No transition or response block is numerically singular. Per-K medians decrease from 1.29×10^{-2} at $K = 2$ to 1.69×10^{-4} at $K = 16$, making the lack of growing-K uniformity visible in the data. The open-density conclusion remains a theorem-level statement rather than an inference from this finite sample.

I.3 Packed storage and update work

The implementation experiment stores each retained grid coordinate as a fixed-length integer index in one packed byte array; only two signed 64-bit counters are stored alongside it. Decoded floating-point arrays are transient workspace and are excluded from persistent-state accounting. We compare this cut–suffix state with an approximate full-window implementation using the same accuracy target, accounting rule, and index representation, and with the numeric payload of the exact full-window cache. The experiment includes $K \in \{2, 3, 4, 8, 16\}$, $\epsilon \in \{2^{-6}, 2^{-8}, 2^{-10}\}$, $D \in \{16, 64, 256\}$, and five seeds; a separate trace study uses $K \in \{2, 3, 8\}$, the same accuracies and delays, four arrival schedules, and three seeds. Every generated prefix is decoded and compared against both the full reference and the earlier floating-point construction.

For the work measurements we run 512 source-time shifts and discard the first 256, so even the $D = 256$ process reaches the measured portion of the stream. The reported study contains 75 storage settings, 108 trace settings, and 159,825 prefix comparisons. Its maximum packed/full TV divided by ϵ is 0.005972, the maximum packed/floating-construction ratio is 0.004825, the largest $D = 256$ packed-byte ratio is 0.05931, and the largest $D = 256$ operation-ratio upper confidence bound is 0.24103. The median wall-time ratio is 3.003, so we make no latency-speedup claim: persistent memory is minimax-matched, while the replay-work saturation is constructive.

Bootstrap intervals resample histories after averaging schedules within a history. Resamples are generated separately for each method–delay setting, so we use them only as marginal intervals,

not for paired differences. The storage comparison charges persistent bytes; transition products quantify algorithmic work; wall time is reported descriptively for the same packed implementation. Configuration files, scripts, aggregate tables, and prefix-level outputs accompany the submission.

J Uniform Bounds for the Reset Family

This appendix proves Corollary 3. All constants introduced below are independent of K, D, ϵ unless an argument is displayed.

J.1 Observation model and contraction

Let $\Omega_h = \{z \in \mathbb{R}^K : \mathbf{1}^\top z = 0, \operatorname{osc}(z) < h\}$ and let ν_h be normalized Lebesgue measure on this bounded open subset of the zero-sum hyperplane. Permuting coordinates preserves Ω_h and ν_h, so

$$Z_h = \int_{\Omega_h} e^{z_s}\, d\nu_h(z)$$

does not depend on s. Therefore $f_s(z) = e^{z_s}/Z_h$ is a legal density, and its reduced log-likelihood field is $(z_1 - z_K, \dots, z_{K-1} - z_K)$. This linear map is an isomorphism from the zero-sum hyperplane to $\mathbb{R}^{K-1}$, so the field image contains an open neighborhood of zero. For the categorical delayed channel, let $t = e^a = 4/3$ and, for outcome $W = j$ under state $X = s$, set

$$\Pr(W = j \mid X = s) = \frac{t^{\mathbf{1}\{s=j\}}}{t + K - 1}.$$

This channel is positive, row-normalized, and contributes oscillation a for every outcome.

For completeness, consider Bayes normalization $M_g(p) = g \odot p/\langle g, p\rangle$. If $m \le g_i \le M$, adding and subtracting $g \odot p'/\langle g, p\rangle$ gives

$$\|M_g(p) - M_g(p')\|_1 \le 2(M/m)\|p - p'\|_1.$$

Rows i, j of P_K differ by $\rho(e_i - e_j)$, hence the Dobrushin coefficient is ρ. A common bounded field has $M/m \le e^{h+a} = 16/9$; the coefficient for Bayes update and prediction, in either order needed by the construction, is at most $2\rho e^{h+a} = 4/9$.

J.2 A quantitative local response

Put $t = e^a = 4/3$, $S = t + K - 1$, and let p be the posterior after outcome one at the uniform fixed point. Then $p_1 = t/S$ and $p_i = 1/S$ for $i > 1$. Since $(P_K^\top p)_i = \rho p_i + (1 - \rho)/K$, differentiating $F_i(r) = \log(P_K^\top p)_i - \log(P_K^\top p)_K$ yields

$$A_K = \begin{pmatrix} \lambda_\parallel & c_K \mathbf{1}^\top \\ 0 & \lambda_\perp I_{K-2} \end{pmatrix}$$

with the three scalars displayed in Section 5. The lower block is absent at $K = 2$, where $|\lambda_\parallel| = 63/25080 > 1/400$.

For $K \ge 3$, $N_K = 24K^2 - 49K - 7$ is positive because $N_3 = 62$ and $N_K' = 48K - 49 > 0$. The polynomial inequality

$$35KN_K - (24K + 7)(25K + 7) = 840K^3 - 2315K^2 - 588K - 49 \ge 0$$

holds at $K = 3$ and increases thereafter, proving $\lambda_\parallel \ge 1/(40K)$. Also $|\lambda_\perp|^{-1} = 8(24K + 7)/7 \le 31K$. The inverse arrow has repeated top-row entry

$$-\frac{c_K}{\lambda_\parallel \lambda_\perp} = -\frac{192K(24K + 7)}{7N_K}.$$

Now N_K/K^2 is minimized at $K = 3$ and $(24K + 7)/K \le 79/3$, so this entry has magnitude below 105. Splitting A_K^{-1} into its block diagonal and off-diagonal row gives

$$\|A_K^{-1}\|_2 \le 40K + 105\sqrt{K - 2} \le 145K < 200K.$$

Thus $\sigma_{\min}(A_K) \ge 1/(200K)$ for every $K \ge 2$.

J.3 PREDICTION GEOMETRY

Fix a logit radius R. If $p = \sigma(r)$ and $\|r\|_\infty \le R$, then for constants depending only on R,

$$c_R/K \le p_i \le C_R/K, \qquad c_R/K \le y_i = (P_K^\top p)_i \le C_R/K. \tag{A1}$$

Write $J_K(r) = DF_K(r)$. From

$$(J_K)_{ij} = \rho\left[\frac{p_i(\mathbf{1}\{i=j\} - p_j)}{y_i} + \frac{p_K p_j}{y_K}\right], \tag{A2}$$

each row and column has one $O_R(1)$ contribution and $K-1$ contributions of order $O_R(1/K)$, so the induced 1- and ∞-norms of J_K are bounded independently of K. We next record the corresponding second-derivative bound explicitly. Put $q = K-1$, $Z = 1 + \sum_{j\le q} e^{r_j}$, and $\beta = (1-\rho)/K$. Then

$$b_i = \rho e^{r_i} + \beta Z \quad (i \le q), \qquad b_K = \rho + \beta Z, \qquad F_i(r) = \log b_i - \log b_K.$$

For directions $u, v \in \mathbb{R}^q$, define

$$s_r(u) = \beta \sum_{j\le q} e^{r_j} u_j, \qquad t_r(u,v) = \beta \sum_{j\le q} e^{r_j} u_j v_j.$$

For $i \le q$,

$$Db_i[u] = \rho e^{r_i} u_i + s_r(u), \qquad D^2 b_i[u,v] = \rho e^{r_i} u_i v_i + t_r(u,v),$$

while the two local terms are absent for b_K. On the R-box, b_i is bounded above and below by positive constants depending only on R, and Cauchy–Schwarz gives

$$|s_r(u)| \le C_R K^{-1/2}\|u\|_2, \qquad |t_r(u,v)| \le C_R K^{-1}\|u\|_2\|v\|_2.$$

The vectors of local terms have norms at most $C_R\|u\|_2$ and $C_R\|u\|_2\|v\|_2$. Applying $D^2 \log b = (D^2 b)/b - (Db \otimes Db)/b^2$ componentwise, the repeated b_K term contributes only $O_R(K^{-1/2})\|u\|_2\|v\|_2$. Consequently there is an $L_R < \infty$, independent of K, such that

$$\sup_{\|r\|_\infty \le R} \|D^2 F_K(r)[u,v]\|_2 \le L_R\|u\|_2\|v\|_2. \tag{A2b}$$

This is the dimension-uniform Hessian estimate used below. The inverse map is the composition

$$r \mapsto y = \sigma(F_K(r)) \mapsto p = \rho^{-1}\{y - (1-\rho)\mathbf{1}/K\} \mapsto \lambda(p).$$

This structure can be bounded without hiding a factor of K. For $s = F_K(r)$, set $a_i = y_i/p_i$. Direct differentiation of the displayed inverse gives, for $i, j \le q$,

$$(J_K^{-1})_{ij} = \rho^{-1}\{a_i \mathbf{1}\{i=j\} + (a_K - a_i)y_j\}. \tag{A3a}$$

Equation (A1) implies $a_i \in [c_R', C_R']$ and $y_j \le C_R/K$. Consequently every row sum is at most $\rho^{-1}(C_R' + C_R' - c_R')$, while every column sum is at most $\rho^{-1}\{C_R' + C_R(C_R' - c_R')\}$. Both bounds are independent of K. After enlarging M_R, therefore $\|J_K^{-1}\|_1, \|J_K^{-1}\|_\infty \le M_R$, and

$$\|J_K\|_2 \le M_R, \qquad \sigma_{\min}(J_K) \ge M_R^{-1}. \tag{A3}$$

Factoring $J_K = D\lambda(y) P_K^\top D\sigma(r)$ on the simplex tangent plane also gives

$$\det J_K(r) = \rho^{K-1} \frac{\prod_{i=1}^K p_i}{\prod_{i=1}^K y_i}. \tag{A4}$$

Equations (A1)–(A4), including (A2b), are the dimension-uniform geometric inputs.

J.4 TRIANGULAR COORDINATES AND PACKING

Retain n consecutive timestamps and use baseline-filter logits $r_0, \dots, r_{n-1}$ as input coordinates. The local fields satisfy $\theta_i = r_i - F_K(r_{i-1})$, so the change from fields is triangular with identity diagonal. Order inputs from newest to oldest. The output is the baseline terminal logit followed by the terminal displacement produced by inserting $v = ae_1$ at each increasing age. Its conditional diagonal block at age e is a product of e prediction Jacobians followed by A_K.

The radius needed below can be chosen uniformly in K. For $\tau \geq 0$, symmetry gives $F_K(\tau e_1) = f_K(\tau)e_1$. Writing $p = \sigma(\tau e_1)$ and $y = P_K^\top p = \rho p + (1-\rho)\mathbf{1}/K$, we have

$$1 \leq \frac{y_1}{y_K} \leq \frac{p_1}{p_K} = e^\tau, \qquad 0 \leq f_K(\tau) \leq \tau.$$

Hence the zero-field baseline and the probe orbit starting from ae_1 lie in $[0, a]e_1$ for every K. The fixed choice $R = a + 1$ contains both and a unit neighborhood of both. Let $M = M_R \geq 2$ in (A2)–(A3), and let $C_2 \geq L_R$ in (A2b). For a composition C_e of e prediction maps whose arguments stay in this box, the chain rule and induction give

$$\|DC_e\|_2 \leq M^e, \qquad \mathrm{Lip}(DC_e) \leq eC_2M^{2e}. \tag{A5}$$

Indeed, the Hessian recurrence is $H_e \leq C_2M^{2(e-1)} + MH_{e-1}$. The local response block is $DF_K(r + v) - DF_K(r)$, so (A2b) also bounds its perturbation by $2C_2\|r - r'\|_2$. Applying the product rule to this block and the subsequent prediction Jacobians yields the same bound in (A5), after enlarging C_2. At the base point the conditional block has smallest singular value at least $M^{-e}/(200K)$. Measure each baseline-block perturbation in Euclidean norm. Perturbing every earlier block by at most δ changes the conditional block, using (A5), the product rule, and a geometric sum, by at most $Cn^2M^{3n}\delta$. Choose $b > 4 \log M$ and then $0 < c \leq (400C)^{-1}$. With

$$\delta_n = \frac{cK^{-1}e^{-bn}}{(n+1)^2} \tag{A6}$$

the perturbation is at most

$$CcK^{-1}e^{-(b-3\log M)n} \leq (400K)^{-1}M^{-n} \leq (400K)^{-1}M^{-e}$$

at every age $e \leq n$. It is therefore no larger than half the base singular value, so each perturbed conditional block has smallest singular value at least $M^{-e}/(400K)$. Taking c smaller also keeps the corresponding local fields inside the open sensor support.

Let the domain be the product of n open Euclidean q-balls of radius δ_n. This fixes the norm used in the perturbation estimate, and every within-block line segment remains in the domain. Integrating each conditional derivative along those segments makes every conditional map co-Lipschitz and one-to-one. Starting from the baseline output and substituting blocks backwards recovers every input, so the full chart is globally one-to-one on the product domain. Multiplying its block determinants and using (A3) gives

$$|\det D\Phi_n| \geq (400K)^{-q(n-1)} \exp(-C_1qn^2). \tag{A7}$$

The response metric is defined in terminal probabilities, not logits. On an R-box the reference-softmax Jacobian is $H = \mathrm{diag}(p_{1:q}) - p_{1:q}p_{1:q}^\top$. For every u,

$$u^\top Hu \geq p_K \sum_{i=1}^{q} p_i u_i^2 \geq c_R K^{-2}\|u\|_2^2. \tag{A8}$$

Strong monotonicity along the segment therefore converts a logit displacement to TV with loss at most C_RK^2. A Euclidean separation in the qn-output chart has one block separated by another factor $\sqrt{n}$. If that block is a terminal-logit displacement, the triangle inequality shows that either its single-probe raw response or the legal empty response is separated by half as much. The product-domain volume is $[\omega_q\delta_n^q]^n$, where ω_q is the unit Euclidean q-ball volume. Stirling's bound gives $-\log \omega_q = O(q \log(q + 1)) = O(q \log K)$. Combining this volume loss, (A7), (A8), the raw-response factor, and a maximal strict packing gives

$$\log \mathcal{P} \geq qnx - C_1qn^2 - C_2qn \log K - C_3qn \log(n + 1). \tag{A9}$$

This displays every state-count loss used in the proof.

Let $n = \min\{D + 1, \max(1, \lfloor \gamma x \rfloor)\}$ with numerical $\gamma > 0$ small enough that the quadratic term in (A9) is at most $qnx/4$. Choose C_0 so $x \geq C_0 \log K$ absorbs the third term and then choose a numerical ϵ_0 so the logarithmic term is also at most $qnx/4$. The response-packing converse proves the lower half of Corollary 3.

For the upper half, before mixing quantize all $D + 1$ bounded field blocks at spacing $\Theta(\epsilon/(D + 1))$. This costs $O(q(D + 1)[x + \log(D + 1)]) = O(q(D + 1)x)$. After mixing retain $L =$

$\lceil \log(3/\epsilon)/|\log(4/9)| \rceil$ edges. Ignoring the older boundary costs at most $\epsilon/3$, and a coordinate-wise field telescope at spacing $\Theta(\epsilon/L)$ costs another $\epsilon/3$. A predicted cut has coordinates in $[(1-\rho)/K, \rho+(1-\rho)/K]$, so its reduced-logit radius is $\log(1+\rho K/(1-\rho))$. Its grid costs $O(qx)$ bits under $x \geq C_0 \log K$; the suffix costs $O(qx^2)$. This proves the upper bound.

For exact dimension take $n = D+1$. The response chart is continuously injective on an open $q(D+1)$-dimensional patch, so invariance of domain gives the converse. Strict positivity makes the node-potential path family minimal; its $q(D+1)$ node marginals are one-to-one mean coordinates and attain the bound.

K Known Inputs and Online Computation

Corollary 4 (Known-input phase law)**.** *Let a revealed word $a_{T-D:T-1}$ select transitions P^{a_t} and both sensor channels, and let $B_\epsilon^{\mathrm{input}}(D)$ be worst-case persistent memory when that word is external. If the model obeys the common bounds of Theorem 2 for every action, then*

$$B_\epsilon^{\mathrm{input}}(D) \leq Cq \min\{D+1, \log(1/\epsilon)\} \log(1/\epsilon)$$

uniformly over input words. If one constant action satisfies Theorem 3 or (OC)*, the lower bound matches and continuous exact dimension is $q(D+1)$.*

Suppose one timestamped packet is generated per source time and receives an independent, state-independent delay in $\{0, \ldots, D\}$ with full support. Arrival reveals the packet's timestamp and value. Let the transport scheduler expose the pending set, and let $B_\epsilon^{\mathrm{corr}}(D)$ charge only persistent correction state.

Theorem 4 (Online correction and system consequences)**.** *For every fixed model in Theorem 3, uniformly over time:*

(a) with the pending set external,

$$B_\epsilon^{\mathrm{corr}}(D) = \Theta(q \min\{D+1, \log(1/\epsilon)\} \log(1/\epsilon))\,;$$

if the filter must recover the exact pending mask, the memory becomes $\Theta(D + B_\epsilon^{\mathrm{corr}}(D))$;

(b) with $L = \min\{D+1, \lceil C \log(1/\epsilon) \rceil\}$, a correction is materialized in $O(T_{\mathrm{mv}}(K)L)$ transition–vector work;

(c) for rewards $r_t(x,a) \in [0,1]$ whose actions do not affect future latent, sensor, or delay laws, the decoded Bayes action has conditional regret at most 2ϵ, H-step regret at most $2H\epsilon$, and discounted regret at most $2\epsilon/(1-\gamma)$.

K.1 Conditioning on revealed inputs

Fix a revealed word $a_{T-D:T-1}$. Conditional on this word, the latent path is a time-varying Markov chain with edge P^{a_t} and action-indexed local fields. The common entry and field bounds are exactly the hypotheses used in Theorem 2: the Doob minorization and its coefficient $\kappa = 1 - m/M$ are uniform in t and in the word. The same projection–round update therefore gives one action-indexed program and a word-independent code length

$$O(q \min\{D+1, \log(1/\epsilon)\} \log(1/\epsilon))\,.$$

Here the revealed input word is external in the same sense as event timestamps. Only the quantized correction state is charged.

For the converse, restrict the supremum over input words to $a_t \equiv a_0$. If this stationary submodel satisfies Theorem 3 or (OC), its response packing is a subset of the responses that a uniform-input code must distinguish, giving the matching lower bound. The same restriction gives the $q(D+1)$ exact-dimensional lower bound. For any fixed revealed word, the $q(D+1)$ node marginals are mean coordinates for the positive node-potential path family and give the exact upper bound. This proves Corollary 4.

K.2 Update work and decision transfer

Let $L_e = \min\{D, \lceil C \log(1/\epsilon) \rceil\}$ be retained edges. During an arrival update the circular buffer contains at most $L_e + 1$ local fields. The persistent finite code is the cut-belief grid index plus these field-grid indices; timestamps are external. Scanning chronologically from the cut uses at most L_e transition-vector products and one transient K-vector. An old packet is rejected by its timestamp in constant time. A time shift absorbs the oldest field into the cut, predicts, projection-rounds, rotates the buffer, and scans at most L_e edges if the output is requested. This proves the work statement in Theorem 4(b).

For a uniform-in-time error check, let h_t be the Hilbert distance between the exact and quantized cut beliefs. A rounded field has log-likelihood oscillation error at most Δ and a rounded cut logit has oscillation error at most Δ_b. The projective argument in Appendix H gives model-dependent constants C_1, C_2 and the Birkhoff coefficient $\tau(P) < 1$ such that

$$h_{t+1} \le \tau(P) h_t + C_1 \Delta + C_2 \Delta_b. \tag{A10}$$

Taking $\Delta = \Theta(\epsilon/(L_e + 1))$ and $\Delta_b = \Theta((1 - \tau(P))\epsilon)$, with model-dependent constants small enough to absorb C_1, C_2, keeps the stationary contribution below the allocated boundary budget. The TV conversion and the ignored-old and retained-field telescopes are exactly those in Appendix H. This proves accuracy after every arrival prefix and time shift. The value $4/9$ is used only for the reset family in Appendix J.

For passive decisions, let $a^\star$ maximize expected reward under b and $\hat{a}$ maximize it under $\hat{b}$. Since every $[0, 1]$-valued expectation changes by at most TV,

$$\mathbb{E}_b r(a^\star) - \mathbb{E}_b r(\hat{a}) \le |\mathbb{E}_b r(a^\star) - \mathbb{E}_{\hat{b}} r(a^\star)| + |\mathbb{E}_{\hat{b}} r(\hat{a}) - \mathbb{E}_b r(\hat{a})| \le 2\epsilon.$$

Action independence lets us sum the conditional inequality over a finite or discounted horizon.

K.3 Scaling study for growing state spaces

The confirmatory grid uses $K \in \{2, 4, 8, 16, 32\}$, $D \in \{16, 32, 64, 128, 256\}$, $\epsilon \in \{2^{-6}, 2^{-8}, 2^{-10}\}$, eight fixed seeds, and 512 shifts. Timing summaries exclude the first 256 shifts so every physical delay has entered the stationary arrival stream. The three implementations are cached full window, cached exact fixed lag, and the quantized cut–suffix method. Method order rotates by seed, time, and prefix. Transition applications are primary; wall time is environment-sensitive.

The resulting tables contain 2,675,610 method–prefix rows, 600 trace settings, and 225 aggregate settings. Direct recomputation of the full belief agrees at all 27,513 checked prefixes. The stored rows cover the complete Cartesian design and support direct checks of accuracy, regret, and transition counts. On-time fields are sampled over the declared support rather than from the conditional sensor distribution. The study therefore probes the theorem's pointwise guarantee; it is not a claim about typical trajectories.

K.4 Learned prediction-versus-revision diagnostic

The learned diagnostic uses a binary reset transition with correlation $\rho = 0.82$. Let F be its one-step logit prediction map. For each terminal logit $s \in [-1, 1]$, we form two length-two histories indexed by $r_0 \in \{-0.9, 0.9\}$:

$$\theta_0 = r_0, \qquad \theta_1 = s - F(r_0).$$

Both histories end at logit s, while adding a late field $v \in \{-1, 1\}$ at the first timestamp changes the terminal logit to $s + F(r_0 + v) - F(r_0)$. Thus every pair is exactly prediction-equivalent and generically revision-distinct. All on-time fields lie in $[-2, 2]$.

We generate 4,500 such pairs and split at the *pair level*: both histories in every pair are assigned together to train, validation, or test. The split contains 3,000/500/1,000 pairs, hence 6,000/1,000/2,000 histories. Every model uses the same one-unit GRU bottleneck and 32-unit tanh heads. We sweep $\lambda \in \{0, .01, .03, .1, .3, 1\}$ in the objective $\ell_{\text{pred}} + \lambda \ell_{\text{corr}}$, with $\lambda = 0$ denoting prediction-only training. After early stopping, we discard the correction head, freeze the encoder, and fit a fresh correction probe with the same architecture and initialization for every condition within a seed. Adam uses learning rate 0.003, batch size 256, and six fixed initialization seeds.

The complete prespecified sweep is shown in Figure 2a. Ordered by increasing λ, mean (forward, revision) MAE pairs in percent are $(.274, 1.250)$, $(.274, 1.240)$, $(.276, 1.220)$, $(.289, 1.154)$, $(.365, .997)$, and $(.650, .680)$. Intervals are two-sided Student-t 95% intervals across seeds ($df = 5$). Every positive weight reduces revision error in all six paired runs. At $\lambda = .1$, forward MAE increases by 0.015 percentage points and revision MAE falls by 7.7%; at $\lambda = 1$, revision MAE falls by 45.6%. The normalized latent gap is the mean absolute distance between prediction-equivalent pair members divided by the test-latent sample standard deviation. It rises from 0.032 to 0.090 between the endpoints. This removes linear scale but not general nonlinear reparameterization, so it is a mechanism diagnostic rather than an invariant representation metric. This controlled witness shows that a forward-trained recurrent state need not preserve correction responses and exposes the cost of preserving them at a fixed bottleneck size. It does not estimate a neural minimax memory law.