How Wrong Can a Good Predictor Be? Diverging Updates with Vanishing Predictive KL

Anonymous authors
Paper under double-blind review

Abstract

Accurate posterior prediction need not require accurate approximation of Bayesian updates. We prove that an unbounded gap between the update maps can coexist with vanishing predictive KL for every fixed finite $K \geq 2$ in a stationary symmetric Gaussian HMM. Exact Bayesian mixing and an explicit deterministic radial filter act on the same $K-1$ belief coordinates. As $q \to 0^+$, their separation in centered logits in the worst case grows at least linearly in the natural confidence scale $L_K(q)$, while their categorical $D_{\mathrm{KL}}(\mathrm{exact}\|\mathrm{radial})$ vanishes at the same explicit witness. Along stationary HMM trajectories, the expected terminal KL between filtered posteriors also converges to zero at $H(q) = \lceil -\log(q)/c \rceil + 1$. Typical blocks without switches drive both filters into a common confidence cone, where softmax curvature suppresses their disagreement; a single Gaussian maximal event controls adaptive noise. A sweep with equally spaced Gaussians over $K \in \{2, 4, 8\}$ illustrates the opposing trends, and binary controls at long horizons compare saturating and nonsaturating recurrences. The result isolates two missing links between internal update gaps and predictive cost: the contribution of separating states to expected loss and decoder sensitivity. Thus even an unbounded internal update gap does not by itself certify predictive failure. The construction is fixed in K and does not provide a universal criterion for when compression is harmless or characterize when internal gaps must incur task loss.

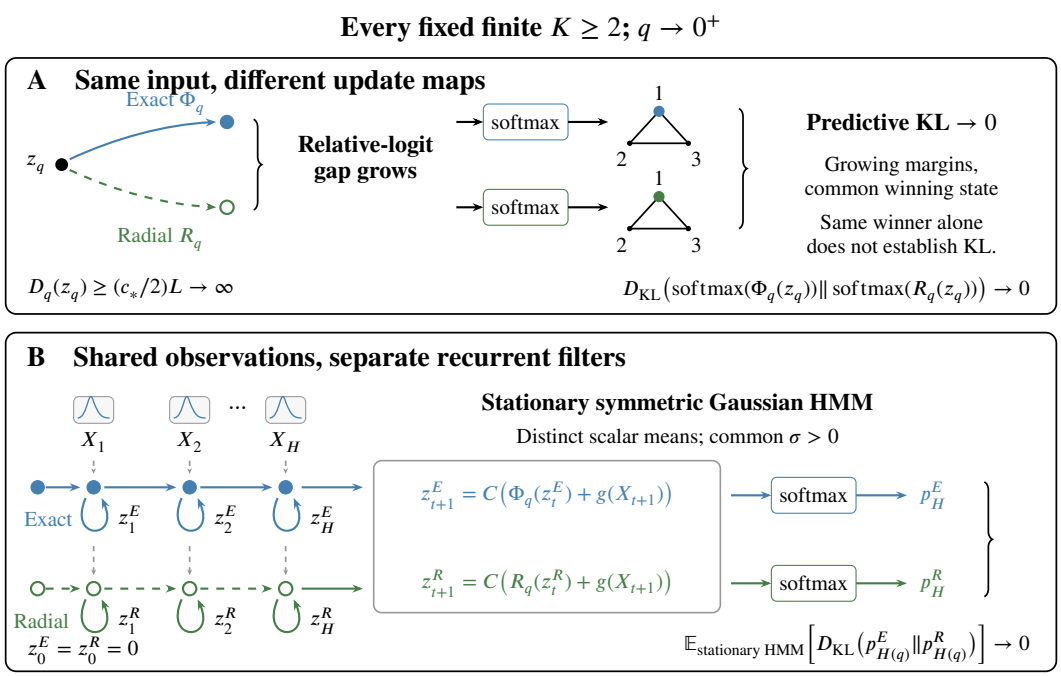
1 Introduction

Transformers retain a growing context, whereas state space and recurrent alternatives compress history into a deterministic state of fixed size. Recurrence expressivity and state tracking are therefore active design constraints (Lahoti et al., 2026; Karuvally et al., 2025), with formal comparisons also separating simplified selective SSMs from linear attention (Cohen-Karlik et al., 2026). But these architectural results do not answer the question at the task level: when does a strict computational mismatch force worse prediction? In a finite HMM, exact Bayes itself occupies only $K-1$ belief coordinates; the issue is whether a different update geometry with a fixed state must pay predictive loss.

Pointwise representability, learnability, and task cost weighted by the distribution are distinct. The predictive cost of a pointwise separation depends on the loss the decoder can see under the evaluation path law, including the loss magnitude on rare paths. Impossibility theorems at the architecture level establish representability gaps in specified models (Shakerinava et al., 2026; Sarrof et al., 2024), whereas learnability can fail even for representable functions (Hahn & Rofin, 2024). Neither distinction alone determines this task cost. We study this boundary in a sequential model with finitely many states where both can be calculated.

Our result has two levels. At the level of the update maps, for every fixed finite $K \geq 2$ an explicit moving witness makes the quotient distance in centered logits between exact Bayesian mixing and a radial filter grow as $\Omega(L_K(q))$, while their categorical $D_{\mathrm{KL}}(\mathrm{exact}\|\mathrm{radial})$ at that same witness tends to zero. At the level of the path law, with pairwise distinct scalar Gaussian means and slow symmetric switching, the filters consume the same stationary HMM observations and their expected

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.



Fix K and emissions before $q \to 0^+$; simplexes show three states.

Figure 1: **Two distinct fixed-K separations.** (A) At a common moving input, exact and radial mixing have diverging separation in centered logits but vanishing categorical KL. (B) On shared stationary observations from the Gaussian HMM, their separate recurrences have vanishing expected terminal posterior KL at $H(q)$. Both statements fix K and emission parameters before $q \to 0^+$; the schematic does not assert that (A) implies (B).

terminal KL between filtered posteriors at $H(q) = \lceil -\log(q)/c \rceil + 1$ also tends to zero. Thus even an unbounded internal separation need not induce a nonvanishing performance gap.

The result at the level of the maps is decoder geometry: both updates become increasingly confident in the same state, so softmax erases a growing logit discrepancy. The theorem under the path law is the main technical contribution, and saturation alone does not give it: the radial filter runs its own recurrence on the shared observations, its effective noise weights depend on its past, and on rare paths the terminal KL is controlled only by an envelope growing like $L_K(q)$, so convergence in probability would not control the expectation. On a block with no latent switch, one maximal event controls all adaptive Gaussian noise weights and radial saturation keeps the approximation in a confidence cone with exact Bayes; the remaining paths have probability $o(1/L_K(q))$ and a uniform KL envelope absorbs them.

A sweep with equally spaced Gaussians over $K \in \{2, 4, 8\}$ illustrates the two opposing trends at finite scale: centered states separate while predictive KL shrinks. The binary case is a microscope rather than the theorem's scope: it visualizes the geometry and separates saturation from tanh specifically. The training experiments do not establish transfer to learned models.

The proof explains how decoder sensitivity and the contribution of separating states to expected loss control predictive error. Controlled numerical experiments illustrate the two results; simpler recurrences also achieve small KL at the tested short horizons (Table 7). We do not claim uniformity in growing K, an impossibility result for a class of architectures, or learnability of the radial rule.

Figure 1 distinguishes the statements at the common input and under the path law.

2 Related Work

Expressivity, learning, and task cost. Efficient sequence models make recurrence expressivity an explicit design variable. Recent work strengthens selective state space updates, analyzes their

limits on tracking the state, and compares simplified selective layers with linear attention (Lahoti et al., 2026; Karuvally et al., 2025; Cohen-Karlik et al., 2026; Shakerinava et al., 2026; Alsmann et al., 2026). Results on the worst case and on formal languages identify computations that specified architectures cannot represent (Merrill et al., 2024; Sarrof et al., 2024), while complementary results show that expressible functions can still be hard to learn (Hahn & Rofin, 2024). These results determine representation or optimization, not the predictive cost weighted by the distribution of a mismatch. Predictive cost additionally depends on decoder sensitivity and the contribution of the separating states weighted by loss under the evaluation path law. With unbounded KL, vanishing event probability alone does not control this contribution. An expressivity analysis for a specific task for GNNs makes a related diagnosis (Kemper et al., 2025); our contribution is an explicit sequential counterexample with a proved loss under the path law.

Harmless information loss and internal divergence. Compression that is aware of the task already shows that discarded information can be harmless relative to a downstream predictor family or distortion (Dubois et al., 2021; 2020; Hafez-Kolahi et al., 2021). We do not optimize a representation or bit rate: the state dimension and prescribed update family are fixed, while their numerical maps depend on q and their distance in centered logits diverges. The generic mathematical shape is not new in isolation. On separable data, gradient descent can send linear predictor norms to infinity while logistic loss tends to zero and the normalized direction converges (Soudry et al., 2018). Our conjunction is different: two explicit filtering maps diverge at the same centered input, their categorical KL vanishes there, and a second theorem proves predictive convergence under a stationary sequential law.

Approximate Bayesian filtering and finite memory. RNNs can approximate Bayesian filtering statistics on fixed finite horizons and, under additional conditions, uniformly in time (Bishop & Bonilla, 2023). Neural Bayesian Filtering instead learns belief embeddings of fixed length and proves particle consistency over a finite horizon under its stated assumptions (Solinas et al., 2025). Prediction with short memory gives a different benchmark: an order-ℓ Markov predictor attains average KL at most $I(\mathcal{M})/\ell$ and average ℓ_1 error at most $\sqrt{I(\mathcal{M})/\ell}$ (Sharan et al., 2018). For an n-state HMM, $I(\mathcal{M}) \leq \log n$, independently of mixing time. This benchmark concerns distributions of the next observation, whereas our theorem measures categorical filtered posteriors. Our predictor maintains $K - 1$ belief coordinates. Streaming projection instead retains a truncated mixture over latent paths with a fixed budget through a deterministic recurrence (Duran-Martin, 2026). The distinction is the maintained object and approximation operation, not recurrence itself; our result couples diverging centered map distance with vanishing categorical KL.

Filter stability and robustness. Classical filter stability asks whether filters with different initial conditions merge while running the correct model (Ocone & Pardoux, 1996; Chigansky et al., 2011; McDonald & Yüksel, 2020). Robustness theory instead perturbs model kernels and controls policy costs (Kara & Yüksel, 2020; Kara et al., 2022) or error in the filter kernel and policy performance (Demirci et al., 2025) under the respective assumptions. Control policies with a finite window can also become nearly optimal as the window grows (Kara & Yüksel, 2022). Our recurrence is neither a differently initialized exact filter nor a sequence of kernels converging to the true one; its pointwise update mismatch grows on the stated witness.

Detection delay and the binary controls. The affine control at the long endpoint is adjacent to quickest change detection. Page introduced the cumulative sum scheme (Page, 1954); Lorden introduced a criterion on conditional delay in the worst case and proved asymptotic optimality, Moustakides established exact CUSUM optimality under it, and Pollak studied nearly minimax detection under a different criterion on conditional delay (Lorden, 1971; Moustakides, 1986; Pollak, 1985). Pollak's irreversible model with a single change is adjacent to but not identical to our HMM that switches repeatedly. These works calibrate the binary diagnostic but do not imply the fixed-K radial theorem.

Why state space models are the motivating case. The design pattern of a deterministic state spans structured SSMs, linear RNNs, gated linear attention, and recurrent competitors (Gu et al., 2020; 2022; Smith et al., 2023; Orvieto et al., 2023; Yang et al., 2024a;b; Peng et al., 2023; Beck et al., 2024; Dao & Gu, 2024). Selective SSM theory and studies of state tracking show

how transition structure, composition, precision, or normalization alter representational capability (Muca Cirone et al., 2024; Siems et al., 2026; Bondaschi et al., 2025; Terzić et al., 2025; Siems et al., 2025; Li & Cotterell, 2025). They motivate the logical question studied here. We do not prove that any named architecture implements, or fails to implement, either filter in our theorem.

3 Problem Setup: Gaussian HMM with a Fixed Finite State and Radial Filter

Definition 1 (Stationary K-state Gaussian HMM and coupled filters) *Fix an integer $K \geq 2$, pairwise distinct means $\mu_0, \ldots, \mu_{K-1} \in \mathbb{R}$, and $\sigma > 0$, with $0 < q < (K-1)/K$ so that $L_K(q) > 0$ and $\alpha_K(q) > 0$. Let $(S_t)_{t\geq 0}$ be stationary and uniform on $\{0, \ldots, K-1\}$, with*

$$\Pr(S_{t+1} = i \mid S_t = i) = 1 - q, \qquad \Pr(S_{t+1} = j \mid S_t = i) = \frac{q}{K-1} \quad (j \neq i).$$

Conditional on $S_t = s$, let $X_{t+1} \sim \mathcal{N}(\mu_s, \sigma^2)$. Write $\mathcal{C}z = z - K^{-1}(\mathbf{1}^\top z)\mathbf{1}$ for centering and

$$g(x)_s = \frac{\mu_s x}{\sigma^2} - \frac{\mu_s^2}{2\sigma^2}$$

for the Gaussian score, understood modulo common shifts. For centered logits z, define exact transition mixing

$$\Phi_q(z) = \mathcal{C}\log(P_q^\top \operatorname{softmax}(z)),$$

where the logarithm is componentwise. Let

$$r(z) = \sqrt{2}\,\|\mathcal{C}z\|_2, \qquad \alpha_K(q) = 1 - \frac{Kq}{K-1}, \qquad L_K(q) = \log\frac{(K-1)(1-q)}{q},$$

and define the radial map

$$R_q(z) = \frac{L_K(q)\tanh(\alpha_K(q)r(z)/L_K(q))}{r(z)}\,\mathcal{C}z,$$

with continuous value $R_q(0) = 0$ and scale $\alpha_K(q)$ at the origin. Starting from $z_0^E = z_0^R = 0$, the coupled filters use the same observations:

$$z_{t+1}^E = \mathcal{C}\big(\Phi_q(z_t^E) + g(X_{t+1})\big), \qquad z_{t+1}^R = \mathcal{C}\big(R_q(z_t^R) + g(X_{t+1})\big).$$

Their terminal categorical outputs are $p_t^E = \operatorname{softmax}(z_t^E)$ and $p_t^R = \operatorname{softmax}(z_t^R)$. For $t \geq 1$, p_t^E is the filtered posterior of S_{t-1} given $X_{1:t}$; p_t^R approximates that posterior. These outputs precede any additional transition or emission channel.

Define

$$d_{\min} = \min_{i\neq j}\frac{(\mu_i - \mu_j)^2}{2\sigma^2}.$$

Fix $0 < c < d_{\min}$ and set $H(q) = \lceil(-\log q)/c\rceil + 1$. Constants below may depend on the fixed K, the full mean vector, σ, and c. Appendix B and Table 4 translate these centered coordinates into a geometric example with three states.

4 Main Result and Proof Mechanism

Definition 2 (Centered separation between the update maps) *For centered $z \in \mathbb{R}^K$, define the quotient distance*

$$\mathcal{D}_q(z) = \max_{i,j}\left|[(\Phi_q(z))_i - (\Phi_q(z))_j] - [(R_q(z))_i - (R_q(z))_j]\right|.$$

It ignores common logit shifts, which do not change a categorical prediction.

Theorem 3 (Diverging updates with vanishing predictive KL) *Fix any finite $K \geq 2$, let $L = L_K(q)$, and set*

$$z_q = L(1/4, -1/4, 0, \ldots, 0), \qquad c_\star = \frac{1}{2} - \tanh\frac{1}{2} > 0.$$

Then

$$\frac{[(\Phi_q(z_q))_0 - (\Phi_q(z_q))_1] - [(R_q(z_q))_0 - (R_q(z_q))_1]}{L} \longrightarrow c_\star \qquad (q \to 0^+).$$

Consequently, for all sufficiently small q, with the threshold allowed to depend on fixed K,

$$\sup_{z:\mathcal{C}z=z} \mathcal{D}_q(z) \geq \mathcal{D}_q(z_q) \geq \frac{c_\star}{2} L_K(q) \longrightarrow \infty.$$

Nevertheless, at the same witness,

$$D_{\mathrm{KL}}(\operatorname{softmax}\Phi_q(z_q) \,\|\, \operatorname{softmax} R_q(z_q)) \longrightarrow 0.$$

Theorem 4 (Predictive convergence with a fixed finite state) ***Setting.*** *For every fixed finite $K \geq 2$, consider Definition 1 with pairwise distinct means, $\sigma > 0$, and $0 < c < d_{\min}$. The two filters share the stationary HMM observation path and are evaluated at $H(q) = \lceil (-\log q)/c \rceil + 1$.*

Conclusion.

$$\mathbb{E}\Big[D_{\mathrm{KL}}\Big(p^E_{H(q)} \,\Big\|\, p^R_{H(q)} \Big) \Big] \longrightarrow 0 \qquad \text{as } q \to 0^+.$$

The expectation is under the stationary HMM path law.

Corollary 5 (Transfer to next-observation prediction) *Let G denote the common Gaussian emission kernel, and let $\mathsf{M}_q = P_q\mathsf{G}$ denote one latent transition followed by emission. For categorical distributions p and r, data processing gives*

$$D_{\mathrm{KL}}(p\mathsf{G} \| r\mathsf{G}) \leq D_{\mathrm{KL}}(p\|r), \qquad D_{\mathrm{KL}}(p\mathsf{M}_q \| r\mathsf{M}_q) \leq D_{\mathrm{KL}}(p\|r).$$

Consequently, the vanishing categorical KL in Theorem 3 implies vanishing KL after the Gaussian emission channel at the same witness. Theorem 4 likewise implies vanishing expected KL between the corresponding next-observation predictive distributions.

Logical separation. Theorem 3 is a statement about the maps in the worst case, with a matched decoder statement: the internal distance diverges at an explicit moving witness, yet the two decoded predictions agree there. Theorem 4 is a distributional statement: predictive agreement also holds in expectation along trajectories generated by the stationary HMM. The first rules out the explanation that the update maps merely become uniformly close; the second shows that the phenomenon is not confined to one logit vector chosen by hand.

Mechanism and proof roadmap. At the explicit witness, both maps place the same coordinate ahead by a margin proportional to $L = L_K(q)$, so categorical softmax curvature suppresses their growing internal discrepancy. Along HMM paths the terminal KL is controlled only by an envelope of order L, so convergence in probability is not enough; the proof bounds the expectation on a good event and on its complement. Condition on the stationary initial state s and let $K \geq 3$. (i) The window contains a latent switch with probability at most $H(q)q$. (ii) Given the state, the centered score is a fixed direction plus one scalar Gaussian noise direction, so the radial state is described by one signal and one noise coefficient (Lemma 8). Because every radial mixing coefficient lies in $(0, 1)$, Abel summation bounds all adaptive noise coefficients by one maximum of a Gaussian partial sum (Lemma 9), which is $O(\sqrt{L \log L})$ outside probability L^{-3}. (iii) A radial barrier and a matching lower bound give every margin of the true state over a wrong state order $L^{2/3}$ after a transient (Lemmas 12 and 14); exact Bayes has margin at least $L/2 - \log K$ outside probability $(K-1)e^{-\kappa L} + CLe^{-L/2}$ (Lemma 17). (iv) In this common confidence cone, softmax curvature gives $D_{\mathrm{KL}} \leq C_K L^2 e^{-c_K L^{2/3}}$ (Lemma 20), while $D_{\mathrm{KL}} \leq 2L$ on every path (Lemma 19). Hence

$$\mathbb{E}\big[D_{\mathrm{KL}}(p^E_H \| p^R_H) \,\big|\, S_0 = s \big] \leq C_K L^2 e^{-c_K L^{2/3}} + 2L\big(Hq + L^{-3} + (K-1)e^{-\kappa L} + CLe^{-L/2} \big) \longrightarrow 0,$$

because $H = \Theta(L)$ and $q \leq (K-1)e^{-L}$. For $K = 2$, Lemma 21 supplies a common logarithmic margin and the same two lemmas conclude. Appendix D explains the exponent $2/3$ and the dependence of the constants on K.

Table 1: **Summary of results.** The first three rows are mathematical results. The fourth is an illustration at finite scale, and the last row records exclusions rather than negative results.

Object	Quantifier	Status
Centered map distance	every fixed finite $K \geq 2$	proved $\Omega(L_K)$
Categorical KL at the same input	every fixed finite $K \geq 2$	proved $\to 0$
Stationary terminal KL	every fixed finite $K \geq 2$	proved $\mathbb{E}\mathrm{KL} \to 0$
Equally spaced Gaussian illustration	$K \in \{2, 4, 8\}$	measured; not used in the proof
Growing K or architecture class	$K = K(q)$; named models	not claimed

Scope. Both results fix K before taking $q \to 0^+$. The quantitative witness moves outward with $L_K(q)$; it does not imply a gap bounded away from zero on every fixed bounded input. Moreover, $H(q)q = O(q \log(1/q)) \to 0$, so Theorem 4 does not characterize post-switch tracking; Appendix F reports a forced-switch diagnostic outside the theorem. Neither theorem is uniform in growing K or nearly colliding means, and neither is an impossibility theorem for an entire SSM architecture class or a claim about what training by gradient descent will learn. Table 1 separates proved statements, evidence at finite scale, and exclusions.

5 Evidence at Finite Scale for Fixed K

Theorem 4 is asymptotic and does not specify a universal finite-q onset. We therefore use a deliberately simple numerical illustration that matches its Gaussian assumptions. For $K \in \{2, 4, 8\}$, the state means are equally spaced, centered, and rescaled to minimum pairwise distance one. We evaluate $L \in \{8, 16, 32, 64, 128\}$ with 4,096 paired stationary paths per cell. Exact Bayes and the radial filter consume the same observations. The two displayed quantities are the centered distance between terminal states and the categorical $D_{\mathrm{KL}}(\mathrm{exact} \| \mathrm{radial})$ after the observation.

Figure 2 shows increasing sampled state distance and decreasing categorical KL across all three displayed values of K: internal distance grows with the scale of rare switching while predictive KL falls. The curves are not used to establish the theorem or to claim a uniform onset in K. Appendix F gives the estimands, coupling, uncertainty calculation, and exact summary statistics.

6 Binary Specialization and Geometry in the Finite Regime

For $K = 2$, translating the two means by their midpoint reduces the model to $(-\mu, +\mu)$ without changing likelihood ratios. Centered logits then collapse to one scalar h. Exact mixing becomes $F_q(h)$ and the radial witness becomes the tanh clamp $G_q(h)$ below. Theorem 3 already supplies an explicit, diverging separation at the level of the maps for every fixed K. This binary slice instead serves as a visualization in the finite regime and isolates why an affine map cannot represent exact mixing; the stationary fixed-K theorem does not depend on the affine obstruction proved here.

Proposition 6 (Geometry of exact mixing) *For every $0 < q < \frac{1}{2}$ the exact mixing map F_q of Definition 1 satisfies:*

(i) ***Saturation****: $|F_q(h)| < L(q)$ for every finite h.*

(ii) ***Contraction****: F_q contracts logit perturbations with global factor at most $1 - 2q$, attained at the midpoint.*

(iii) ***Not affine****: with $x = \log 2$ and*

$$\Delta_q := \left|F_q(0) - 2F_q(x) + F_q(2x)\right| > 0,$$

every affine map A satisfies $\max_{z \in \{0, x, 2x\}} |F_q(z) - A(z)| \geq \Delta_q/4$.

Proof. See Appendix E. □

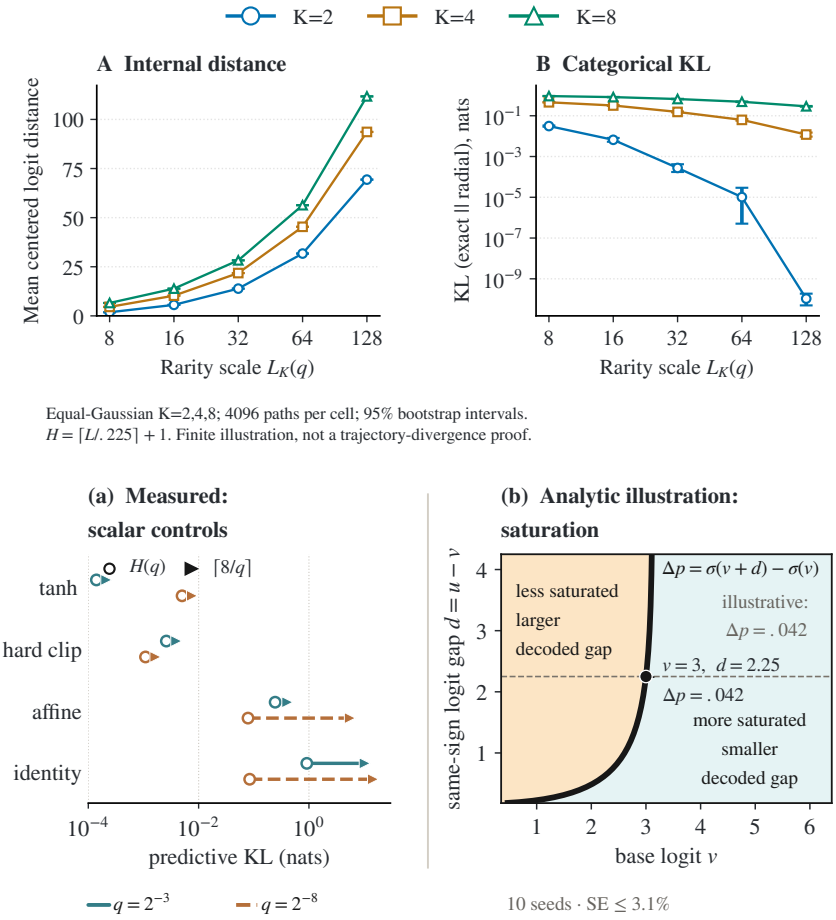


Figure 2: **Separation at finite scale and decoder saturation.** Top: centered distance between terminal states grows while categorical $D_{\mathrm{KL}}(\mathrm{exact}\|\mathrm{radial})$ falls for Gaussian HMMs with $K = 2, 4, 8$ (4,096 paired paths per cell; 95% bootstrap intervals). Bottom: KL of the binary controls over ten seeds at $H(q)$ and $\lceil 8/q \rceil$ (relative SE $\leq 3.1\%$), followed by an analytic illustration with a logistic decoder, not sampled trajectories. Nearly equal endpoints are displaced for legibility. Table 7 and Appendix F give values, clipping conventions, and estimation details.

Remark 7 (The obstruction does not give a lower bound) *Part (iii) of Proposition 6 is adverse evidence for a naive pointwise argument. Direct expansion gives* $\Delta_q = \frac{3}{4}q + O(q^2)$ *as* $q \to 0^+$. *Thus the obstruction is strict for every* $q > 0$ *but its quantitative margin is not bounded away from zero along the family with rare switching, so it cannot yield a nonvanishing asymptotic predictive lower bound.*

Moreover, predictive loss is evaluated after the logistic decoder under the HMM path law, not in the logit norm in the worst case. The approximate map G_q preserves the exact saturation scale and midpoint slope, and the shared envelope $|F_q(h)|, |G_q(h)| < L(q)$ gives global control without requiring uniform closeness.

The exact, radial, and affine transition geometry is shown in Figure 3 in the appendix.

7 Binary Controls and Learned Transfer

Three comparisons each answer one question: whether learned recurrences reach the proved mechanism at the theorem's horizon (Table 2), whether that match survives at $\lceil 8/q \rceil$, and whether saturation, rather than tanh specifically, explains the binary behavior (scalar controls). The experiment with learned models asks whether recurrences faithful to the architecture recover the proved

Table 2: **Closure per q when trained end to end, mean $\pm$ standard error over ten evaluation seeds.** Closure places the constrained scalar S6 arm at 0 and the analytic tanh recurrence at 1, so the arm faithful to the architecture meets the criterion of half closure only at $q \in \{2^{-5}, 2^{-6}, 2^{-7}, 2^{-8}\}$, the four empirically passing points. The asymptotic theorem supplies no finite-q cutoff.

q	**Mamba** $\times 2$	**GRU**	**Finite window**
Below empirical half closure			
2^{-3}	-1.59 ± 0.39	0.61 ± 0.06	-2.75 ± 0.65
2^{-4}	-0.09 ± 0.21	1.00 ± 0.03	-1.17 ± 0.47
Above empirical half closure			
2^{-5}	1.04 ± 0.02	1.11 ± 0.03	0.04 ± 0.20
2^{-6}	1.14 ± 0.01	1.18 ± 0.01	0.68 ± 0.06
2^{-7}	1.13 ± 0.02	1.14 ± 0.02	0.74 ± 0.07
2^{-8}	1.16 ± 0.01	1.19 ± 0.01	0.87 ± 0.03

mechanism across six switch probabilities. When trained end to end, the two-block Mamba arm (Mamba $\times 2$ in Table 2) closes at least half the reference gap at four of six q values, below the prespecified criterion of five out of six; these results do not establish transfer to learned models. Under distillation, every nonreference comparison arm meets the threshold of half closure at all six values.

Closure is defined as $(\mathrm{KL}_{\mathrm{S6}} - \mathrm{KL}_{\mathrm{arm}})/(\mathrm{KL}_{\mathrm{S6}} - \mathrm{KL}_{\tanh})$, so the constrained scalar S6 arm sits at 0, the analytic tanh recurrence sits at 1, and values above 1 mean the learned arm beats the proved mechanism. Every architecture loss is $D_{\mathrm{KL}}(\text{exact Bayes} \| \text{model})$ after float32 probability clipping, unlike the 10^{-15} clipping of the scalar controls in Table 7; Appendix F states both conventions. Table 2 gives the breakdown at the horizon matched to the theorem $H(q) = \lceil -\ln q \rceil + 1$.

The learned arm passes at the four smaller tested q values, but this empirical split does not identify the theorem's asymptotic onset. At $q = 2^{-3}$ the reference gap $\mathrm{KL}_{\mathrm{S6}} - \mathrm{KL}_{\tanh}$ is only 1.5×10^{-3}, so closure is scale sensitive and seed dispersion is widest at the two largest q. Even so, the two points below half closure lie 5.4 and 2.8 standard errors below the threshold, whereas the four passing points lie at least 26 standard errors above it using the unrounded results across seeds. The outcome of four out of six is therefore not explained by marginal uncertainty. Under distillation, the closure computed from the ratio of mean losses ranges from 1.16 to 3.97 and every nonreference comparison arm passes at all six q values. Table 2 instead reports the mean of closure ratios for each seed.

Generalization at long horizons separates the learned architectures. The secondary endpoint on long paths evaluates the same trained models on sequences of length $\lceil 8/q \rceil$, up to 2048 steps, against a training horizon matched to the theorem that is at most eight. Here the ranking inverts. The analytic tanh recurrence is horizon stable, its predictive KL staying between 1.7×10^{-4} and 6.4×10^{-3} at the two tested switch probabilities, while the scalar S6 arm degrades to 1.9×10^{-1}. The GRU inherits that stability, passing four of six q when trained end to end and all six under distillation. The official Mamba blocks do not: when trained end to end, two stacked blocks pass only two of six, and a single block fails at every q, with a worst per-q mean closure at long horizons of -32.65 at $q = 2^{-3}$ when trained end to end. Under distillation, both Mamba depths pass zero of six at this long endpoint. We therefore do not claim that a canonical selective SSM acquires the mechanism. Matching the exact filter at the theorem's horizon does not ensure stability at longer horizons. Per-seed long-horizon losses and closures for every arm are in the supplement (`empirical/data/ten-seed-results.csv`).

Scalar controls test a simpler explanation. A naive recurrence could also achieve low loss at the horizon matched to the theorem, making the exhibited tanh geometry unnecessary. We therefore measured three controls against exact Bayes with the same generator ($\mu = \sigma = 1$, 200,000 paths per point at $H(q)$, 20,000 at $\lceil 8/q \rceil$, split across ten fixed seeds). They are a plain affine recurrence $h \leftarrow (1 - 2q)h + \ell(x)$ with no saturation, the class considered by the pointwise lower bound; a hard clip of that map at $\pm L(q)$; and the identity accumulator. The third panel of Figure 2 shows the endpoint pattern; Appendix F, Table 7, reports the exact cells. Each recurrence is scored after its own transition map, so these cells measure one-step predictive states rather than the filtered posteriors of Theorem 4.

The controls support three conclusions. First, both saturating recurrences are horizon stable at the two tested switch probabilities, while the affine and identity controls degrade at the long endpoint. Their relative ordering is not uniform: the hard clip has lower predictive KL at $q = 2^{-8}$, whereas the tanh clamp has lower predictive KL at $q = 2^{-3}$. These two points do not identify a crossover threshold. The theorem should therefore be read as proving one saturating instance, not the optimal one. Second, the endpoint matched to the theorem does not establish a separation among scalar recurrence classes; the binary empirical separation among these scalar controls occurs at the long endpoint and is measured, not proved. At that endpoint the affine map reaches 5.53 nats at $q = 2^{-8}$, while both saturating controls remain stable. Third, the closure scale of Table 2 anchors 1 to one saturating witness among several; the readings from the learned arm remain meaningful as calibrated distances to an analytic reference with a guarantee at the stated logarithmic horizon. Its stability at longer horizons is measured here, not established by that theorem.

Hyperparameters were selected by mean validation loss over three tuning seeds before confirmation, with ties broken deterministically and incomplete settings excluded. Tuning and confirmation seeds are disjoint, all final evaluation seeds are included, and the supplement gives the procedure for model selection and implementation details.

8 Limitations

Both theorems fix finite K before taking $q \to 0^+$. Their constants can deteriorate with K and with nearly colliding means; no rate uniform in $K = K(q)$ is proved. The quantitative witness moves with $L_K(q)$, so it gives neither a gap on every bounded input nor a lower bound on the state dimension. The stationary theorem further assumes pairwise distinct scalar Gaussian means, symmetric switching, stationary initialization, and the stated logarithmic horizon.

The sweep with equally spaced Gaussians is an illustration at finite scale, not evidence for a uniform onset rate. The comparison concerns two explicit maps, not every SSM, and does not show that training learns the radial rule. Binary controls are diagnostics in the finite regime: hard clipping can beat tanh, and the training experiments do not establish transfer to learned models. The result is a counterexample, not a universal criterion for pruning, quantization, distillation, low rank adaptation, or other compression. The converse question, when an internal mismatch must incur nonvanishing task loss, remains open.

9 Conclusion

For every fixed finite latent state space, exact Bayesian mixing and our radial filter can diverge linearly in centered logits at an explicit witness while their decoded categorical KL vanishes. Predictive agreement also holds in stationary expectation in the symmetric scalar Gaussian HMM, so it is neither explained by uniform internal approximation nor confined to one input chosen by hand. A controlled sweep with equally spaced Gaussians over $K \in \{2, 4, 8\}$ illustrates both trends at finite scale, while binary controls compare behavior at short and long horizons. At the common input witness, softmax suppresses a growing map discrepancy; on typical HMM paths, a shared confidence cone suppresses predictive error. The result therefore identifies two missing links from internal separation to predictive loss: its contribution weighted by loss under the evaluation path law, and the decoder's sensitivity there. It does not show that this failure of gap transfer is prevalent in learned models or provide a converse criterion; uniform growing-K rates and extensions to learned sequence models remain open.

Reproducibility statement

The assumptions and complete paper proofs for the stated results are given in the appendices. Appendix F specifies the empirical estimands, aggregation, model configurations, hyperparameters, seeds, and clipping conventions. The anonymous supplementary package contains the simulation code and numerical results used for the reported experiments, mapped in its `README.md`: the complete original Gaussian grid (`empirical/fixed-k-bridge/results/full/`) and its separately labeled extension (`empirical/fixed-k-bridge/results/k8-extension/`), all 600 learned-model evaluations and the 120 per-seed rows at both endpoints (`empirical/data/`

records.json, empirical/data/ten-seed-results.csv), scalar-control raw losses (empirical/scalar-controls.replay/), and the forced-switch diagnostic (results/forced-switch-recovery-2026-09-21/).

AI USE DISCLOSURE

Generative AI tools assisted with literature search and organization, manuscript drafting and editing. The authors checked the cited sources, reviewed all AI-assisted text. The authors take responsibility for the final content of this work, including text and artifacts produced with the aid of generative AI.

REFERENCES

Eric Alsmann, Lowejatan Noori, and Martin Lange. On the expressiveness of state space models via temporal logics. In *International Conference on Learning Representations*, 2026.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, Dylan Zinsley, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff. *arXiv preprint arXiv:2402.18668*, 2024.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xLSTM: Extended long short-term memory. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2405.04517.

Philipp Becker, Niklas Freymuth, and Gerhard Neumann. KalMamba: Towards efficient probabilistic state space models for RL under uncertainty. *arXiv preprint arXiv:2406.15131*, 2024.

Adrian N. Bishop and Edwin V. Bonilla. Recurrent neural networks and universal approximation of bayesian filters. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6956–6967. PMLR, 2023. URL https://proceedings.mlr.press/v206/bishop23a.html.

Marco Bondaschi, Nived Rajaraman, Xiuying Wei, Kannan Ramchandran, Razvan Pascanu, Caglar Gulcehre, Michael Gastpar, and Ashok Vardhan Makkuva. From markov to laplace: How mamba in-context learns markov chains, 2025. arXiv:2502.10178.

Loïc Cabannes, Pierre-Emmanuel Mazaré, Gergely Szilvasy, Matthijs Douze, Maria Lomeli, Ilze Amanda Auzina, Justin Carpentier, Gabriel Synnaeve, and Hervé Jégou. Sparse delta memory: Scaling the state of linear RNNs through sparsity. *arXiv preprint arXiv:2607.07386*, 2026.

Pavel Chigansky, Robert Liptser, and Ramon van Handel. Intrinsic methods in filter stability. In *The Oxford Handbook of Nonlinear Filtering*, pp. 319–351. Oxford University Press, 2011.

Edo Cohen-Karlik, Itamar Zimerman, Liane Galanti, Ido Andrew Atad, Amir Globerson, and Lior Wolf. On the expressivity of selective state-space layers: A multivariate polynomial approach. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 136–144. PMLR, 2026. URL https://proceedings.mlr.press/v300/cohen-karlik26a.html.

Adrien Corenflos, James Thornton, George Deligiannidis, and Arnaud Doucet. Differentiable particle filtering via entropy-regularized optimal transport. In *International Conference on Machine Learning*, 2021. arXiv:2102.07850.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, 2024. arXiv:2405.21060.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. Neural networks and the chomsky hierarchy. In *International Conference on Learning Representations*, 2023. arXiv:2207.02098.

Yunus Emre Demirci, Ali Devran Kara, and Serdar Yüksel. Sensitivity of filter kernels and robustness bounds to transition and measurement kernel perturbations in partially observable stochastic control. *arXiv preprint arXiv:2508.10658*, 2025.

Yann Dubois, Douwe Kiela, David J. Schwab, and Ramakrishna Vedantam. Learning optimal representations with the decodable information bottleneck. In *Advances in Neural Information Processing Systems*, 2020. arXiv:2009.12789.

Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, and Chris J. Maddison. Lossy compression for lossless prediction. In *Advances in Neural Information Processing Systems*, 2021. arXiv:2106.10800.

Gerardo Duran-Martin. A predictive view on streaming hidden markov models, 2026. arXiv:2604.09208.

Meir Feder and Yury Polyanskiy. Sequential prediction under log-loss and misspecification. *arXiv preprint arXiv:2102.00050*, 2021.

Daniel Y. Fu, Tri Dao, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. In *International Conference on Learning Representations*, 2023. arXiv:2212.14052.

Aydin Ghojogh, M. Hadi Sepanj, and Benyamin Ghojogh. On the relation of state space models and hidden markov models. *arXiv preprint arXiv:2601.13357*, 2026.

Tingnan Gong, Junghwan Lee, Xiuyuan Cheng, and Yao Xie. Neural network-based CUSUM for online change-point detection. *arXiv preprint arXiv:2210.17312*, 2022.

Riccardo Grazzi, Julien Siems, Arber Zela, Jörg K. H. Franke, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear RNNs through negative eigenvalues. *arXiv preprint arXiv:2411.12537*, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. arXiv:2312.00752.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. HiPPO: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems*, 2020. arXiv:2008.07669.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. arXiv:2111.00396.

Hassan Hafez-Kolahi, Behrad Moniri, Shohreh Kasaei, and Mahdieh Soleymani Baghshah. Rate-distortion analysis of minimum excess risk in bayesian learning. In *International Conference on Machine Learning*, 2021. arXiv:2105.04180.

Michael Hahn and Mark Rofin. Why are sensitive functions hard for transformers? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. arXiv:2402.09963.

Ramin Hasani, Mathias Lechner, Tsun-Hsuan Wang, Makram Chahine, Alexander Amini, and Daniela Rus. Liquid structural state-space models. In *International Conference on Learning Representations*, 2023. arXiv:2209.12951.

Vishesh Jain, Frederic Koehler, Jingbo Liu, and Elchanan Mossel. Accuracy-memory tradeoffs and phase transitions in belief propagation. *arXiv preprint arXiv:1905.10031*, 2019.

Sham M. Kakade, Akshay Krishnamurthy, Gaurav Mahajan, and Cyril Zhang. Learning hidden markov models using conditional samples. *arXiv preprint arXiv:2302.14753*, 2023.

Ali Devran Kara and Serdar Yüksel. Robustness to incorrect system models in stochastic control. *SIAM Journal on Control and Optimization*, 2020. arXiv:1803.06046.

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.

Ali Devran Kara, Maxim Raginsky, and Serdar Yüksel. Robustness to incorrect models and data-driven learning in average-cost optimal stochastic control. *Automatica*, 2022. arXiv:2003.05769.

Arjun Karuvally, Franz Nowak, T. Anderson Keller, Carmen Amo Alonso, Terrence J. Sejnowski, and Hava T. Siegelmann. Bridging expressivity and scalability with adaptive unitary SSMs. *arXiv preprint arXiv:2507.05238*, 2025.

Niklas Kemper, Tom Wollschläger, and Stephan Günnemann. What expressivity theory misses: Message passing complexity for GNNs. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2509.01254.

Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations*, 2026. arXiv:2603.15569.

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), 2004. doi: 10.1214/aoap/1075828050.

Jiaoda Li and Ryan Cotterell. Characterizing the expressivity of fixed-precision transformer language models. *arXiv preprint arXiv:2505.23623*, 2025.

Marten Lienen, Abdullah Saydemir, and Stephan Günnemann. UnHiPPO: Uncertainty-aware initialization for state space models. *arXiv preprint arXiv:2506.05065*, 2025.

G. Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6), 1971. doi: 10.1214/aoms/1177693055.

Curtis McDonald and Serdar Yüksel. Exponential filter stability via dobrushin's coefficient. *Electronic Communications in Probability*, 2020. arXiv:1910.08463.

Arian Mehrfard, Bharanidhar Duraisamy, Stefan Haag, Florian Geiss, and Mirko Mählisch. Adaptive learned state estimation based on KalmanNet. *arXiv preprint arXiv:2604.02441*, 2026.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *International Conference on Machine Learning*, 2024.

George V. Moustakides. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4), 1986. doi: 10.1214/aos/1176350164.

Nicola Muca Cirone, Antonio Orvieto, Benjamin Walker, Cristopher Salvi, and Terry Lyons. Theoretical foundations of deep selective state-space models. In *Advances in Neural Information Processing Systems*, 2024.

Daniel Ocone and Etienne Pardoux. Asymptotic stability of the optimal filter with respect to its initial condition. *SIAM Journal on Control and Optimization*, 34(1):226–243, 1996. doi: 10.1137/S0363012993256617.

Antonio Orvieto, Samuel L. Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning*, 2023. arXiv:2303.06349.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1-2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Xingjian Du, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Jiaju Lin, Przemysław Kazienko, Jan Kocoń, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Bolun Wang, Johan S. Wind, Stanisław Woźniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. RWKV: Reinventing RNNs for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

Liangzu Peng, Aditya Chattopadhyay, Luca Zancato, Elvis Nunez, Wei Xia, and Stefano Soatto. Gated KalmaNet: A fading memory layer through test-time ridge regression. *arXiv preprint arXiv:2511.21016*, 2025.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, 2023. arXiv:2302.10866.

Moshe Pollak. Optimal detection of a change in distribution. *The Annals of Statistics*, 13(1), 1985. doi: 10.1214/aos/1176346587.

Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Yash Sarrof, Yana Veitsman, and Michael Hahn. The expressive capacity of state space models: A formal language perspective. In *Advances in Neural Information Processing Systems*, 2024.

Vaisakh Shaj, Cameron Barker, Aidan Scannell, Andras Szecsenyi, Elliot J. Crowley, and Amos Storkey. Kalman linear attention: Parallel bayesian filtering for efficient language modelling and state tracking. In *International Conference on Machine Learning*, 2026.

Mehran Shakerinava, Behnoush Khavari, Siamak Ravanbakhsh, and Sarath Chandar. The expressive limits of diagonal ssms for state-tracking. In *International Conference on Learning Representations*, 2026.

Vatsal Sharan, Sham Kakade, Percy Liang, and Gregory Valiant. Prediction with a short memory, 2018. arXiv:1612.02526.

Julien Siems, Timur Carstensen, Arber Zela, Frank Hutter, Massimiliano Pontil, and Riccardo Grazzi. DeltaProduct: Improving state-tracking in linear RNNs via householder products. *arXiv preprint arXiv:2502.10297*, 2025.

Julien Siems, Riccardo Grazzi, Korbinian Pöppel, Kirill Kalinin, Hitesh Ballani, and Babak Rahmani. Learning state-tracking from code using linear RNNs, 2026. arXiv:2602.14814.

Jimmy T.H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations*, 2023. arXiv:2208.04933.

Christopher Solinas, Radovan Haluška, David Sychrovský, Finbarr Timbers, Nolan Bard, Michael Buro, Martin Schmid, Nathan R. Sturtevant, and Michael Bowling. Neural bayesian filtering, 2025. arXiv:2510.03614.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, and Nathan Srebro. The implicit bias of gradient descent on separable data. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=r1q7n9gAb`.

Alex Tang, M. Emrullah Ildiz, Batin Kurt, Samet Oymak, and Necmiye Ozay. On the generalization properties of selective state-space models for filtering tasks for unknown systems. *arXiv preprint arXiv:2604.23818*, 2026.

Aleksandar Terzić, Nicolas Menet, Michael Hersche, Thomas Hofmann, and Abbas Rahimi. Structured sparse transition matrices to enable state tracking in state-space models. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2509.22284.

Benjamin Walker and Terry Lyons. Chess-world-model: A 10M-game benchmark for exact state tracking from chess move sequences. *arXiv preprint arXiv:2605.30100*, 2026.

Geoffrey Wolfer and Aryeh Kontorovich. Estimating the mixing time of ergodic markov chains. *arXiv preprint arXiv:1902.01224*, 2019.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *International Conference on Machine Learning*, 2024a. arXiv:2312.06635.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems*, 2024b. arXiv:2406.06484.

Di Zhang and Jiaqi Xing. Bayesian optimality of in-context learning with selective state spaces, 2026. arXiv:2602.17744.

A Open Affine Control Lower Bound

This appendix concerns a different question from Theorem 4: whether the nonsaturating binary affine control has a nonvanishing lower bound at the much longer endpoint $\lceil 8/q \rceil$. The section on learned models establishes that separation only by measurement; the final analytic step for the joint event below remains open. The mechanism is *sluggishness*: after a latent switch, the nonsaturating affine recurrence carries an unclamped state near its AR(1) fixed point $d/(2q)$, with $d = 2\mu^2/\sigma^2$, and needs $\Theta(1/q)$ steps to change sign, while exact Bayes recovers on a shorter timescale for building confidence. Simulation localizes the cost accordingly: the fraction on the wrong side relative to the truth is between 0.18 and 0.20 across every q tested. The subset on which Exact Bayes is confident is reported separately below. Disagreement with the exact filter was not measured. Table 3 separates the proved ingredients from the remaining step for the terminal joint event. For the concentration calculation, the centered oriented evidence has variance proxy $v = 2d$, and at horizon t we choose the deterministic prefix budget $B_t = td$.

Table 3: **The lower bound per step, step by step.** Every deterministic and probabilistic ingredient is proved; the banded final row is the one step still open, and the separation at long horizons is not claimed without it.

Step	Statement	Status
Deterministic core		
Cost of a step on the wrong side	a confident exact predictive against a model on the wrong side costs $\geq 3/80$ nats	proved
Affine state	$h_t = a^t h_0 + \sum_{i<t} a^{t-1-i} e_i$	proved
Summation by parts	$\lvert S_j \rvert \leq B_t$ gives weighted sum $\geq -2B_t$	proved
Combining the bounds	no crossing by step t, and that step costs $\geq 3/80$	conditional on the margin
Probabilistic tail		
Gaussian sub-Gaussianity	a centered real Gaussian is sub-Gaussian with constant its variance	proved here
Prefix tail on a block	on a block with fixed latent signs, some prefix exceeds B_t with measure at most $2(t{+}1)e^{-B_t^2/(2tv)}$	proved, discharges the tail
Terminal joint event	**uniform probability that the terminal step is on the wrong side while exact Bayes is confident**	**open**

With $v = 2d$ and $B_t = td$, the charge from the prefix tail is $2(t{+}1)e^{-td/4}$. Instantiating the affine recurrence at $a = 1 - 2q$ and $h_0 = d/(2q)$ reduces the condition for no crossing $2B_t < a^t h_0$ to

$$4qt < (1 - 2q)^t.$$

Writing $y = qt$ and taking $q \to 0$ gives the boundary equation $4y = e^{-2y}$, whose unique positive root is $y_\star = 0.175866\ldots$. Hence $t^* q \to y_\star$, and stationarity gives limiting mass of the run age $1 - e^{-y_\star} = 0.161270\ldots$. Both constants are analytic, independent of q, and carry no sampling error. The latter is close to the measured fraction on the wrong side relative to the truth, 0.18 to 0.20 across $q = 2^{-5}$ to 2^{-12} under the convention after mixing of Table 2; the bound from summation by parts loses a factor of two, so the analytic constant need not match the observed fraction.

What the open step actually requires. The estimand at the long endpoint is the predictive KL at the *final step*, so no integration along a trajectory is needed: it suffices to lower bound, uniformly in q, the probability that the terminal step is a step on the wrong side, since the cost per step above then gives $\mathbb{E}[\mathrm{KL}_T] \geq (3/80)\,\mathbb{P}(W_{\mathrm{joint}})$. Stationarity supplies that bridge exactly. Writing A_T for the age of the current latent run, memorylessness gives $\mathbb{P}(A_T \geq k) = (1 - q)^k$, hence $\mathbb{P}(A_T \leq t^*) \to 1 - e^{-0.176}$, which is the 0.161 already reported. What remains open is therefore a uniform bound on the joint event that the run is young enough for the affine state not to have crossed and old enough for the exact filter to be confident, not an integration.

Why a stepwise union bound is insufficient. Charging the worst case at every step forces the tolerance per step to grow like $\sqrt{\log(1/q)}$, shrinking the guaranteed horizon before crossing to $\Theta(1/(q\sqrt{\log(1/q)}))$ and making the resulting probability of the switch window tend to zero. This decay conflicts with the measured nearly constant fraction on the wrong side. Controlling the partial sum instead preserves a constant probability of the switch window.

B Geometry of Centered Logits and an Example with Three States

This section gives a geometric reading of Definition 1. It is not an additional assumption and is not used as a substitute for the proof. Its purpose is to make clear what the state dimension, centering operation, and radial nonlinearity mean when $K > 2$.

B.1 What K counts

The integer K counts latent modes, not observations, layers, or time steps. A machine with normal, overheated, and failed modes has $K = 3$ even if it is monitored for a million steps. At time t, a filter assigns a probability vector

$$p_t = (p_t(0), \dots, p_t(K-1)), \qquad p_t(s) \geq 0, \quad \sum_s p_t(s) = 1.$$

The vector has K entries but only $K - 1$ degrees of freedom because its entries sum to one. Logits make the same redundancy explicit: for every scalar b,

$$\mathrm{softmax}(z + b\mathbf{1}) = \mathrm{softmax}(z).$$

Thus z and $z + b\mathbf{1}$ represent the same belief. Centering chooses one representative from each equivalence class,

$$\mathcal{C}z = z - \frac{\mathbf{1}^\top z}{K}\mathbf{1}, \qquad \mathbf{1}^\top \mathcal{C}z = 0.$$

The meaningful filter state therefore lies in the $(K-1)$-dimensional hyperplane orthogonal to $\mathbf{1}$. Pairwise log odds $z_i - z_j$ are coordinates on this quotient space. This is why every mismatch in Definition 2 is stated through pairwise differences rather than raw logits.

It is useful to keep two scale parameters separate. The state count K fixes the dimension of the belief geometry. The quantity

$$L_K(q) = \log \frac{(K-1)(1-q)}{q}$$

is a confidence scale determined by the switch probability. For fixed K, rare switching means $q \to 0^+$ and hence $L_K(q) \to \infty$. The theorem takes this limit after fixing K; it does not let the number of modes grow with rarity.

B.2 Direction records preference; radius records confidence

Within the centered hyperplane, regard z as an arrow. Its direction records the pattern of relative preferences among states. Its length records how strongly the filter prefers that pattern. Our convention $r(z) = \sqrt{2}\|\mathcal{C}z\|_2$ makes this length agree with the absolute log odds in the binary specialization.

For a concrete $K = 3$ belief, take

$$z = (2, 0, -2), \qquad \mathrm{softmax}(z) \approx (0.867, 0.117, 0.016).$$

State 0 leads state 1, which leads state 2. If a radial operation halves the arrow, it produces $(1, 0, -1)$ and probabilities approximately $(0.665, 0.245, 0.090)$. The ranking and all ratios between pairwise gaps are unchanged, but the confidence is lower. This example illustrates the geometry only; the actual radial factor is chosen by q and by the current radius.

More precisely, the radial transition can be written

$$R_q(z) = \beta_q(z)\mathcal{C}z, \qquad \beta_q(z) = \frac{L_K(q)\tanh(\alpha_K(q)r(z)/L_K(q))}{r(z)}.$$

Table 4: **Geometric dictionary for the fixed-K result.** Observation scores are added after the transition map and may rotate the state.

Object	Operational meaning
K	Number of possible hidden modes; the centered belief state has dimension $K-1$.
q and $L_K(q)$	Switch rarity and its associated confidence scale; $L_K(q)$ grows as switches become rarer.
$\mathcal{C}z$	Logit state after removing the common offset that does not affect prediction.
$r(z)$	Overall confidence magnitude, not a state count and not a time horizon.
R_q	Nonlinear confidence cap that preserves centered direction during transition mixing; it is not a linear projection.

It preserves the direction of every nonzero centered state during the transition mixing step and changes only its radius. It is not an orthogonal projection: there is no fixed subspace onto which the state is dropped, and the multiplier depends nonlinearly on the current radius. Near the origin, $R_q(z) \approx \alpha_K(q)\mathcal{C}z$; far from the origin, its radius saturates below $L_K(q)$. The next observation then adds the centered score $\mathcal{C}g(x)$, which can change both direction and length. A trajectory can therefore turn many times even though each isolated radial mixing step preserves direction.

B.3 What Exact Bayes does differently

Exact transition mixing acts on probabilities before returning to centered logits. It adjusts the coordinates according to the full categorical mixture $P_q^\top p$, not through one shared radial multiplier. Except in special directions, it need not preserve the centered arrow's direction. The radial filter deliberately discards this geometry specific to each coordinate and retains only a common confidence contraction followed by the same observation score.

The two main theorems test different consequences of that difference. Theorem 3 selects an explicit moving logit z_q and proves that the exact and radial maps separate by order $L_K(q)$ in centered coordinates. This rules out the explanation that the two update maps merely converge to one another. Theorem 4 then couples both filters to the same stationary observation path and proves that their decoded predictive distributions nevertheless converge in expected KL at the logarithmic horizon.

The apparent paradox is resolved by the softmax decoder. A centered displacement is expensive near a decision boundary, where probabilities are sensitive to logits. It can be cheap deep inside a shared confidence cone, where both filters put almost all mass on the same state. In the example with three states, moving from $(10, 0, -10)$ to $(6, 0, -6)$ is a large internal change, yet both decoded vectors are overwhelmingly concentrated on state 0. The proof formalizes precisely this regime inside the common cone rather than asserting that every large logit difference is harmless.

B.4 A dictionary for one step

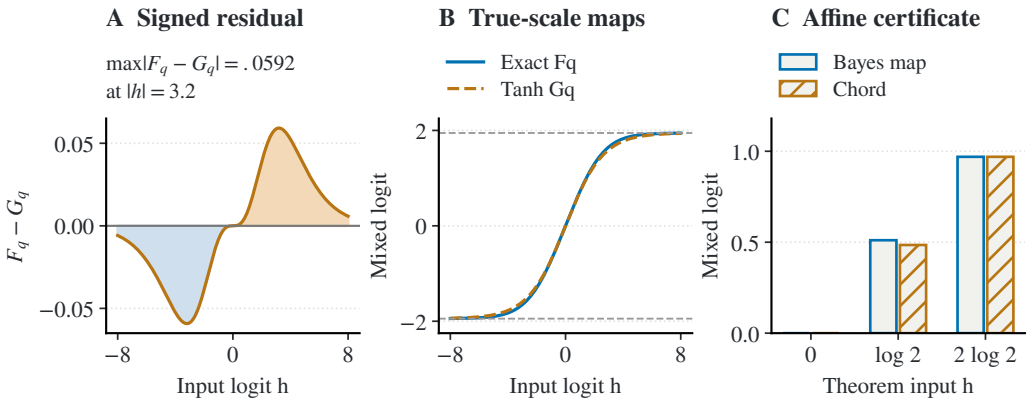
The following dictionary separates objects that are easy to conflate.

C Proof Roadmap for Fixed K

This section is a reading guide to Appendix E. The quantitative theorem at the level of the maps and the stationary theorem under the path law share an endpoint set by softmax curvature but answer different questions. The first follows from an explicit centered witness and direct asymptotics. The second has a probabilistic spine and two recovery arguments specific to the dimension.

Dependency order. Theorem 3 first compares the exact and radial gaps for the chosen pair at z_q, then invokes the curvature lemma for the common cone. For Theorem 4 and $K \geq 3$, Lemma 9 and the Gaussian maximal event feed the radial barriers and Lemma 14; separately, the calculation over path mixtures yields Lemma 17. These two margins meet only in Lemma 20, with the determin-

istic envelope controlling the complement. The binary branch replaces the radial barrier step with common scalar recovery and then rejoins the same curvature and envelope argument.



Fixed $q = 1/8$; full 1601-point grid evaluated.
Chord midpoint error: $\Delta_q/2 = .02613$. Any affine map: max error $\geq \Delta_q/4 = .01306$.

Figure 3: **Binary transition geometry at $q = 1/8$.** A: signed residual of exact minus tanh. B: the maps at true scale share saturation levels and midpoint slope. C: values at $0, \log 2, 2\log 2$ certify that the map is not affine; their second difference Δ_q forces every affine approximation to incur maximum error at least $\Delta_q/4$. This is a fixed-q, $K = 2$ diagnostic.

C.1 Proof outline

Both filters start in the same centered belief coordinates and consume the same observations. The proof under the path law first conditions on a constant latent block. It then shows that exact Bayes and the radial recurrence identify the true state with diverging margins. A deterministic lemma on the categorical KL converts that common confidence cone into an exponentially small loss on the good event, while a $2L_K(q)$ envelope controls every remaining path.

C.2 The $K \geq 3$ branch

On a block with no switch in state s, the centered Gaussian score is $a_s + wZ_t$. Because radial mixing uses one adaptive scalar coefficient, the approximate state has the exact representation $z_t^R = U_t a_s + V_t w$. Abel summation bounds every V_t by twice one maximum of a Gaussian partial sum, so a single maximal inequality controls the whole horizon without a temporal union bound. Projection onto $w^\perp$ supplies a deterministic signal direction; upper and lower radial barriers then give margins for the true state of order $L_K(q)^{2/3}$. Recovery of the exact filter is established separately by comparing the constant path in the true state with constant wrong paths and the aggregate Bayes factor of all alternatives that contain a switch.

C.3 The binary branch

For $K = 2$, the signal and noise directions are collinear, so the projection step is unavailable. Midpoint translation reduces arbitrary distinct means to the symmetric model. An event on the prefix evidence forces both scalar recurrences to hit a common margin, and a geometric sum over possible hit locations controls suffix persistence. Figure 3 visualizes the transition maps used by this branch. The argument unions suffix lengths, not transition errors at one step, and its failure probability remains negligible after multiplication by the KL envelope.

C.4 Conditioning argument

There are three changes of probability law. First, the event that no switch occurs costs at most $H(q)q$. Second, conditional on that event, Gaussian concentration applies to independent emissions from the fixed latent state. Third, the exact posterior still sums over every latent path; candidate

paths that contain a switch are included and controlled through their aggregate Bayes factor. The final expectation is recovered by averaging the finitely many conditional bounds over the stationary uniform starting state.

D Proof Anatomy, Constants, and Quantifiers

This section explains the main steps in the proof of Theorem 4 and the dependence of its constants on the model parameters. Full proofs are given in Appendix E.

D.1 Quantifier order

Fix K, the distinct means $\mu_0, \dots, \mu_{K-1}$, the variance σ^2, and a horizon constant $c < d_{\min}$, where

$$d_{\min} = \min_{i \neq j} \frac{(\mu_i - \mu_j)^2}{2\sigma^2} > 0.$$

Only then send $q \to 0^+$. Constants denoted C_K, c_s, κ, or a small-q threshold may depend on these fixed parameters. This permits a finite union over states and competitors, but it does not provide a bound uniform in K or in a family whose means approach one another. The theorem has the logical form

$$\forall K < \infty \; \forall (\mu_0, \dots, \mu_{K-1}) \text{ distinct}, \qquad \lim_{q \to 0^+} \mathbb{E} D_{\mathrm{KL}}(p_H^E \| p_H^R) = 0,$$

not a joint limit over K and q.

D.2 The probabilistic argument for $K \geq 3$

Condition on a block with no switch whose true state is s. The centered Gaussian score has the exact form with two vectors

$$\mathcal{C}g(X_{t+1}) = a_s + wZ_{t+1},$$

so the radial recurrence stays in the random plane spanned by a_s and w:

$$z_t^R = U_t a_s + V_t w, \qquad U_{t+1} = \beta_t U_t + 1, \quad V_{t+1} = \beta_t V_t + Z_{t+1}.$$

This does not reduce the filter to two states. It reduces the analysis of one conditioned block to one accumulated signal coefficient U_t and one accumulated noise coefficient V_t.

The multiplier β_t depends on the past, so the noise is not a deterministic weighted Gaussian sum. The key observation is order, not independence: for fixed terminal t, the products of positive multipliers attached to later observations are nondecreasing. Abel summation therefore gives the pathwise inequality

$$|V_t| \leq 2 \max_{u \leq H} \left| \sum_{k=1}^{u} Z_k \right|$$

simultaneously for all t. One maximal bound from an exponential martingale controls this maximum. The proof does not apply a separate tail bound at every time and then union bound over the horizon.

On the resulting event, $|V_t| = O(\sqrt{L \log L})$. The tanh expansion supplies an upper barrier $U_t = O(L^{2/3})$ and a matching lower bound after an initial transient of order $L^{2/3}$. The exponent $2/3$ is the balance at which the unit signal increment competes with the cubic correction in $\tanh x = x - \Theta(x^3)$. It is sufficient for recovery because

$$z_{t,s}^R - z_{t,j}^R = d_{sj} U_t + \eta_{sj} V_t, \qquad d_{sj} = \frac{(\mu_s - \mu_j)^2}{2\sigma^2} > 0, \quad \eta_{sj} = \frac{\mu_s - \mu_j}{\sigma},$$

and $L^{2/3}$ dominates $\sqrt{L \log L}$.

Exact Bayes is handled separately. The constant-s latent path is the likelihood reference. Constant wrong paths lose exponentially in L, while the aggregate Bayes factor of all switched paths has expectation of order Hq. This is a calculation over path mixtures; it does not assume independence between posterior coordinates. Because $H = \Theta(L)$ and $q = \Theta(e^{-L})$ for fixed K, mass of the switched paths vanishes.

D.3 Bounding predictive KL

The radial and exact arguments meet only after both filters have entered a common confidence cone. On the good event, every margin of the true state over a wrong state is at least $c_* L^{2/3}$. Along the line segment between the two terminal logits, softmax curvature is then exponentially small. Combining this curvature with the deterministic terminal range bound yields

$$D_{\mathrm{KL}}(p_H^E \| p_H^R) \le C_K L^2 e^{-c_* L^{2/3}}.$$

The bound on this event alone is not enough because KL is unbounded in arbitrary logits. Lemma 19 supplies a pathwise $2L$ envelope, which is multiplied by the vanishing probability of the complement. Combining the bound on this event with the uniform bound on its complement turns recovery with high probability into convergence of expected KL.

D.4 Why the binary branch is separate

For $K \ge 3$, distinct scalar means imply that a_s cannot be collinear with the common noise direction w: otherwise a nonzero quadratic would agree with an affine function at at least three distinct points. This makes the radial radius control available. With two points, that polynomial contradiction disappears. The binary proof instead translates arbitrary means to the symmetric pair $(-\delta, +\delta)$ and controls both scalar recurrences through a common recovery event. It then rejoins the same KL curvature and envelope lemmas. The separate branch is an issue of proof geometry, not a restriction to symmetric models with two states.

D.5 Rates proved and rates not claimed

The witness at the level of the maps and stationary path result have different rates and should not be merged. At the explicit witness, centered map separation is asymptotically $c_\star L$, and decoder KL is bounded by a polynomial in L times $e^{-L/5}$. Along typical blocks with no switch in the stationary proof, the radial margin for the true state is only required to grow as $L^{2/3}$, which still makes the decoder cost on the good event vanish. Neither statement identifies an optimal exponent, a sharp finite-q onset, or constants uniform over state geometries. The curves with equally spaced Gaussians in Appendix F are therefore illustrations rather than rate estimates.

E Proofs of the Main Results

This appendix proves Theorems 3 and 4 for every fixed finite $K \ge 2$. The quantitative theorem at the level of the maps uses an explicit centered witness and a direct asymptotic calculation. The stationary theorem uses a representation in two scalars of the radial filter, one Gaussian maximal event, and an analysis over path mixtures of exact Bayes for $K \ge 3$; the binary case uses the same KL bounds after midpoint translation and a recovery argument in one dimension.

E.1 Preliminaries

Write $T = -\log q$, $\alpha = \alpha_K(q)$, and $L = L_K(q)$. For fixed K,

$$L = T + \log(K-1) + \log(1-q) = T + O(1).$$

Consequently, for all sufficiently small q, $L > 1$, $\alpha \ge 1/2$, and $H \le C_H L$. We will also use

$$q = \frac{K-1}{e^L + K - 1} \le (K-1)e^{-L}, \qquad Hq \longrightarrow 0, \qquad (1-q)^{-(H-1)} \longrightarrow 1. \tag{10}$$

Fix a state s and let

$$\mathcal{N}_s = \{S_0 = \cdots = S_{H-1} = s\}.$$

Conditional on $\mathcal{N}_s$, $X_{t+1} = \mu_s + \sigma Z_{t+1}$ for $t = 0, \dots, H-1$, where the Z_t are independent standard Gaussians. With

$$a_s = \mathcal{C}\left(\frac{\mu_s \boldsymbol{\mu}}{\sigma^2} - \frac{\boldsymbol{\mu}^{\odot 2}}{2\sigma^2}\right), \qquad w = \mathcal{C}\left(\frac{\boldsymbol{\mu}}{\sigma}\right), \tag{11}$$

the centered score is $g(X_{t+1}) = a_s + w Z_{t+1}$.

Lemma 8 (Radial dynamics in two scalars) *On $\mathcal{N}_s$, adapted scalar processes U_t, V_t satisfy*

$$z_t^R = U_t a_s + V_t w, \qquad U_0 = V_0 = 0,$$
$$U_{t+1} = \beta_t U_t + 1, \qquad V_{t+1} = \beta_t V_t + Z_{t+1}, \tag{12}$$

where

$$\beta_t = \frac{L \tanh(\alpha r(z_t^R)/L)}{r(z_t^R)}$$

with continuous value α at the origin. Moreover, $0 < \beta_t \le \alpha < 1$ and $U_t \ge 0$.

Proof. The radial transition multiplies every centered coordinate by the same scalar β_t. Substitution of (11) proves (12) by induction. Positivity is immediate, while $\tanh x \le x$ gives $\beta_t \le \alpha$. □

E.2 One maximal event controls the adaptive noise

Let $S_u = \sum_{k=1}^{u} Z_k$.

Lemma 9 (Adaptive Abel bound) *For every realized path and every $t \le H$,*

$$|V_t| \le 2 \max_{u \le H} |S_u|. \tag{13}$$

Proof. Expanding (12) gives

$$V_t = \sum_{k=1}^{t} \gamma_{k,t} Z_k, \qquad \gamma_{k,t} = \prod_{\ell=k}^{t-1} \beta_\ell, \qquad \gamma_{t,t} = 1.$$

Although the weights are adaptive, they obey $0 \le \gamma_{1,t} \le \cdots \le \gamma_{t,t} = 1$ pathwise. Abel summation therefore gives

$$V_t = S_t - \sum_{k=1}^{t-1} (\gamma_{k+1,t} - \gamma_{k,t}) S_k.$$

The coefficients in the second term are nonnegative and sum to at most one, proving (13). This pathwise bound also applies to adaptive weights. □

Lemma 10 (Gaussian maximal event) *Set*

$$x_L = \sqrt{2H \log(2L^3)}, \qquad \mathcal{E}_L = \left\{ \max_{u \le H} |S_u| \le x_L \right\}.$$

Then $\Pr(\mathcal{E}_L^c) \le L^{-3}$. On $\mathcal{E}_L$,

$$\max_{t \le H} |V_t| \le M_L := 2x_L = O(\sqrt{L \log L}). \tag{14}$$

Proof. For every $\lambda > 0$, $\exp(\lambda S_u - u\lambda^2/2)$ is a nonnegative martingale. Doob's maximal inequality, optimized at $\lambda = x/H$, yields

$$\Pr\left(\max_{u \le H} S_u > x \right) \le e^{-x^2/(2H)}.$$

Apply the same argument to $-S_u$ and union only the two signs. Substituting $x = x_L$ and applying Lemma 9 proves the claim simultaneously over the full horizon. □

E.3 Radial recovery when $K \ge 3$

Assume $K \ge 3$, and let P be Euclidean projection onto $w^\perp$.

Lemma 11 (Noncollinear signal) *For every state s, $Pa_s \ne 0$.*

Proof. Distinct means imply $w \neq 0$. If $Pa_s = 0$, centeredness would give $a_s = \lambda w$. Hence a constant b would satisfy

$$\frac{\mu_s \mu_i}{\sigma^2} - \frac{\mu_i^2}{2\sigma^2} = \frac{\lambda \mu_i}{\sigma} + b$$

for every i. After clearing denominators, a quadratic with nonzero quadratic coefficient would vanish at the $K \geq 3$ distinct values μ_i, a contradiction. □

Set $D_s = \sqrt{2}\|Pa_s\|_2 > 0$, $A_s = r(a_s)$, and $W = r(w)$. From Lemma 8,

$$D_s U_t \leq r(z_t^R) \leq A_s U_t + W|V_t|. \tag{15}$$

Lemma 12 (Upper radial barrier) *For each s, constants $C_U, C_r, q_s > 0$ exist such that, on $\mathcal{E}_L$ and for $q < q_s$,*

$$U_t \leq C_U L^{2/3}, \qquad r(z_t^R) \leq C_r L^{2/3} \tag{16}$$

for all $t \leq H$.

Proof. The function $h(x) = \tanh(x)/x$, continuously extended at zero, decreases on $[0, \infty)$. Using the lower bound in (15),

$$\beta_t U_t = \alpha h(\alpha r(z_t^R)/L) U_t \leq \frac{L}{D_s} \tanh(\alpha D_s U_t / L).$$

Let $f_L(u) = 1 + (L/D_s)\tanh(\alpha D_s u/L)$. Choose C_U so that $D_s^2 C_U^3 > 64$, and put $R = C_U L^{2/3}$. For large L, $\alpha D_s R/L \leq 1$. Since $\tanh x \leq x - x^3/8$ on $[0, 1]$,

$$f_L(R) - R \leq 1 - (1-\alpha)R - \frac{\alpha^3 D_s^2 C_U^3}{8} \leq 1 - \frac{D_s^2 C_U^3}{64} < 0.$$

The map f_L is increasing, $U_0 \leq R$, and $U_{t+1} \leq f_L(U_t)$, so induction on the first crossing yields $U_t \leq R$. The upper bound in (15), together with (14) and $M_L = o(L^{2/3})$, gives the radius bound. □

Lemma 13 (Lower memory after the initial transient) *There are constants $c_U, C_0, q_s' > 0$ such that, on $\mathcal{E}_L$ and for $q < q_s'$,*

$$U_t \geq c_U L^{2/3} \tag{17}$$

whenever $C_0 L^{2/3} \leq t \leq H$.

Proof. The global inequality $\tanh x \geq x - x^3/3$ and Lemma 12 imply

$$\beta_t \geq \alpha - \frac{\alpha^3 r(z_t^R)^2}{3L^2} \geq 1 - \varepsilon_L, \qquad \varepsilon_L = \frac{Kq}{K-1} + \frac{C_r^2}{3L^{2/3}}. \tag{18}$$

For small q, $0 < \varepsilon_L < 1$ and $b_1 L^{-2/3} \leq \varepsilon_L \leq b_2 L^{-2/3}$ for positive constants b_1, b_2. Iterating (12) gives

$$U_t \geq \frac{1 - (1-\varepsilon_L)^t}{\varepsilon_L}.$$

For $t \geq \lceil (\log 2)/\varepsilon_L \rceil$, the numerator is at least $1/2$, which proves (17). This transient is $O(L^{2/3})$ and is smaller than $H = \Theta(L)$. □

Lemma 14 (Simultaneous radial recovery) *For every s, constants $c_s, q_s'' > 0$ exist such that, on $\mathcal{E}_L$ and for $q < q_s''$,*

$$z_{t,s}^R - z_{t,j}^R \geq c_s L^{2/3} \tag{19}$$

for every $j \neq s$ and $C_0 L^{2/3} \leq t \leq H$.

Proof. Direct calculation gives

$$(a_s)_s - (a_s)_j = d_{sj} := \frac{(\mu_s - \mu_j)^2}{2\sigma^2} > 0, \qquad w_s - w_j = \eta_{sj} := \frac{\mu_s - \mu_j}{\sigma}.$$

Thus $z_{t,s}^R - z_{t,j}^R = d_{sj} U_t + \eta_{sj} V_t$. The first term is bounded below by a positive multiple of $L^{2/3}$, uniformly over the finitely many competitors, whereas the second is $O(\sqrt{L \log L}) = o(L^{2/3})$ by (14). □

E.4 Recovery of the exact filter by path mixtures

Continue to condition on $\mathcal{N}_s$. The exact posterior after H observations is the normalized sum of prior times likelihood over latent paths. Use the constant path $(s, \dots, s)$ as reference; its prior mass is

$$\pi_* = \frac{1}{K}(1-q)^{H-1}. \tag{20}$$

Lemma 15 (Constant wrong paths) *There are $\kappa, q_s > 0$ such that, for $q < q_s$, with conditional probability at least $1-(K-1)e^{-\kappa L}$, every constant path $j \neq s$ has likelihood ratio relative to the reference at most $e^{-L/2}$.*

Proof. For $j \neq s$, let

$$Y_t^{s,j} = \log \frac{\varphi_{\mu_s,\sigma}(X_t)}{\varphi_{\mu_j,\sigma}(X_t)}.$$

Under $\mathcal{N}_s$, these variables are independent Gaussians with mean d_{sj} and variance $2d_{sj}$. Since $Hd_{sj} \geq (d_{\min}/c)T$ with $d_{\min}/c > 1$, while $L = T + O(1)$, a bound on the Gaussian lower tail gives

$$\Pr\left(\sum_{t=1}^{H} Y_t^{s,j} < L/2\right) \leq e^{-\kappa_{sj} L}$$

for small q. A finite union over $j \neq s$ proves the claim. □

Lemma 16 (Bayes factor of a switched path) *Let $\mathcal{P}_{\mathrm{sw}}$ be the latent paths containing at least one switch, let $\pi(p)$ denote their prior masses, and let LR_p be likelihood relative to the constant-s path. For*

$$B_{\mathrm{sw}} = \frac{\sum_{p\in\mathcal{P}_{\mathrm{sw}}} \pi(p)\mathrm{LR}_p}{\pi_*},$$

$$\mathbb{E}[B_{\mathrm{sw}} \mid \mathcal{N}_s] \leq \frac{KHq}{(1-q)^{H-1}}, \qquad \Pr(B_{\mathrm{sw}} > e^{-L/2} \mid \mathcal{N}_s) \leq CLe^{-L/2}. \tag{21}$$

Proof. Under the reference observation law, every likelihood ratio has expectation one. The numerator in the first expectation is therefore the prior mass of paths that contain a switch, at most Hq. Division by (20) gives the first bound. Markov's inequality at threshold $e^{-L/2}$, followed by (10), gives the second. □

Lemma 17 (Exact posterior margin) *Conditional on $\mathcal{N}_s$, outside an event of probability at most*

$$(K-1)e^{-\kappa L} + CLe^{-L/2}, \tag{22}$$

the exact terminal logits satisfy

$$z_{H,s}^E - z_{H,j}^E \geq L/2 - \log K \tag{23}$$

for every $j \neq s$.

Proof. On the intersection of the events in Lemmas 15 and 16, total nonreference weight divided by reference weight is at most $Ke^{-L/2}$. Hence $p_H^E(s)/p_H^E(j) \geq e^{L/2}/K$ for every $j \neq s$, and taking logarithms gives (23). □

E.5 Categorical KL bounds

For vectors u, v, put $\Delta = u - v$ and $\mathrm{range}(\Delta) = \max_i \Delta_i - \min_i \Delta_i$.

Lemma 18 (Range envelope) *For every $u, v \in \mathbb{R}^K$,*

$$D_{\mathrm{KL}}(\mathrm{softmax}\, u \| \,\mathrm{softmax}\, v) \leq \mathrm{range}(u - v). \tag{24}$$

Proof. Let $A(z) = \log \sum_i e^{z_i}$. The increment $A(u) - A(v)$ lies between $\min_i \Delta_i$ and $\max_i \Delta_i$. Thus every log probability ratio is at most $\mathrm{range}(\Delta)$, and averaging under $\mathrm{softmax}(u)$ proves (24). □

Lemma 19 (Terminal envelope) *For every observation path,*

$$D_{\mathrm{KL}}(\operatorname{softmax} z_H^E \| \operatorname{softmax} z_H^R) \le 2L. \tag{25}$$

Proof. Let $m^E = \Phi_q(z_{H-1}^E)$ and $m^R = R_q(z_{H-1}^R)$ be the logits before the observation at the terminal step. The final shared score cancels, so $z_H^E - z_H^R = m^E - m^R$. Exact mixing places every probability coordinate in $[q/(K-1), 1-q]$, hence every exact pairwise score before the observation difference has magnitude at most L. The radial image has pairwise diameter strictly below L because its radial norm is below L. Therefore $\operatorname{range}(m^E - m^R) \le 2L$, and (24) applies. □

Lemma 20 (Curvature on a common cone) *Suppose a state s and $m \ge 0$ satisfy*

$$u_s - u_j \ge m, \qquad v_s - v_j \ge m$$

for every $j \ne s$, and $\operatorname{range}(u - v) \le R$. *Then*

$$D_{\mathrm{KL}}(\operatorname{softmax} u \| \operatorname{softmax} v) \le K(K-1)R^2 e^{-m}. \tag{26}$$

Proof. Every point $u_\tau = (1-\tau)u + \tau v$ has the same margin lower bound. If $p_\tau = \operatorname{softmax}(u_\tau)$, then $1 - p_\tau(s) \le (K-1)e^{-m}$. The Hessian of log-sum-exp is $\operatorname{diag}(p_\tau) - p_\tau p_\tau^\top$, so its operator norm is at most $2(K-1)e^{-m}$. The Hessian annihilates common shifts, and $\operatorname{range}(u - v) \le R$ implies $\|\mathcal{C}(u - v)\|_2^2 \le KR^2$. The integral Taylor formula for the corresponding Bregman divergence now gives (26). □

E.6 Binary specialization

When $K = 2$, let

$$m_0 = \frac{\mu_0 + \mu_1}{2}, \qquad \delta = \frac{|\mu_1 - \mu_0|}{2}.$$

After relabeling if needed and translating observations by m_0, the emission means are $-\delta, +\delta$. This preserves the likelihood ratio, the latent law, and both recurrences for the pairwise logits. Their evidence drift and variance are

$$d = \frac{2\delta^2}{\sigma^2} = \frac{(\mu_1 - \mu_0)^2}{2\sigma^2} = d_{\min}, \qquad \operatorname{Var}(Y_t) = 2d. \tag{27}$$

In pairwise log odds the two transition maps are

$$F_q(h) = 2\operatorname{artanh}((1-2q)\tanh(h/2)), \qquad G_q(h) = L\tanh((1-2q)h/L). \tag{28}$$

Lemma 21 (Binary common recovery) *Let $r = \lceil T/c \rceil$, $H = r + 1$, $a = 3\log T$, and $A = B = 5a$. There are events $\mathcal{B}_q$ such that*

$$\Pr(\mathcal{B}_q) L^2 \longrightarrow 0, \tag{29}$$

and on $\mathcal{B}_q^c$ the exact and radial posterior log odds at time H, oriented toward the block's initial state, are both at least a.

Proof. For $x \ge 0$, define the common budget for one step for the confidence loss

$$\rho_q(x) = \max\left\{\log(1 - q + qe^x) - \log(1-q),\ 2qx + \frac{x^2}{L}\right\}. \tag{30}$$

The two entries bound the losses of F_q and G_q, respectively, on a positive interval. Conditional on a block with no switch and its initial sign, the signed evidence variables $Y_0, \dots, Y_r$ are independent $\mathcal{N}(d, 2d)$. Put $N = r + 1$ and

$$Q = L + A + r\rho_q(A), \qquad \Sigma_{\mathrm{pre}} = \exp\left[-\frac{(Nd - Q)^2}{4Nd}\right]. \tag{31}$$

If $\sum_{i=0}^r Y_i > Q$, both recurrences must hit A: otherwise the lower bounds for one step $F_q(x) \ge x - \rho_q(A)$ and $G_q(x) \ge x - \rho_q(A)$ sum from the uniform reset to a terminal value strictly above A, a contradiction.

After a hit, monotonicity and saturation give, for either map M_q,

$$M_q(x) \geq \min\{x, B\} - \rho_q(B). \tag{32}$$

Thus the hit persists to margin a whenever every nonempty suffix obeys

$$\sum_{i=j}^{r} \big(Y_i - \rho_q(B)\big) > -(A - a). \tag{33}$$

For a suffix of length ℓ, the bound on the Gaussian lower tail is at most

$$\exp\left[-\frac{\{A - a + \ell[d - \rho_q(B)]\}^2}{4\ell d}\right].$$

Summing over possible hit locations, and using $\rho_q(5a) \to 0$, we have eventually $d - \rho_q(B) \geq d/2$. With $A - a = 4a$,

$$\frac{(4a + \ell d/2)^2}{4\ell d} \geq a + \frac{\ell d}{16}.$$

The entire suffix sum is therefore at most $C_d e^{-a} = C_d T^{-3}$.

For the prefix, choose $d_0 = (c + d)/2$. Eventually $Q \leq N d_0$, and

$$L^2 \Sigma_{\text{pre}} \leq T^2 \exp\left[-\frac{r(d - c)^2}{16d}\right] \longrightarrow 0.$$

Removing the conditioning on no switch separately from the prefix and suffix bounds costs at most $2Nq$, and $2NqL^2 = O(T^3 e^{-T}) \to 0$. Combining these three terms proves (29) and the deterministic hit and persist argument proves the common margin. □

E.7 Completion of the main theorem

First let $K \geq 3$. Conditional on $S_0 = s$, intersect the actual event of no switch $\mathcal{N}_s$, the maximal event $\mathcal{E}_L$, and the two events for exact recovery from Lemmas 15 and 16. On this good event, Lemmas 14 and 17 give a common terminal margin for the true state at least $c_* L^{2/3}$ for some $c_* > 0$. Lemmas 19 and 20 then yield

$$D_{\text{KL}}(p_H^E \| p_H^R) \leq C_K L^2 e^{-c_* L^{2/3}}. \tag{34}$$

The complement has conditional probability at most

$$Hq + L^{-3} + (K - 1)e^{-\kappa L} + CLe^{-L/2}. \tag{35}$$

This unions event types only; temporal control is already contained in the single maximal event. Applying the deterministic $2L$ envelope (25) on the complement shows that the conditional expected KL tends to zero. The stationary starting state is uniform, so averaging the finitely many conditional bounds over s proves Theorem 4 for $K \geq 3$.

For $K = 2$, Lemma 21 and (26), with $R = 2L$, give on $\mathcal{B}_q^c$

$$D_{\text{KL}}(p_H^E \| p_H^R) \leq CL^2 e^{-3 \log T} \longrightarrow 0.$$

On $\mathcal{B}_q$, (25) and (29) give $\mathbb{E}[D_{\text{KL}}; \mathcal{B}_q] \leq 2L \Pr(\mathcal{B}_q) \to 0$. This proves the binary case for arbitrary distinct means and completes the proof of Theorem 4.

E.8 Proof of diverging updates with vanishing predictive KL

Proof of Theorem 3. Fix finite $K \geq 2$ and abbreviate

$$\alpha = 1 - \frac{Kq}{K - 1}, \qquad L = L_K(q), \qquad b = \frac{q}{K - 1}.$$

Adding a common constant to the input leaves $\text{softmax}(z)$, and therefore $\log(P_q^\top \text{softmax}(z))$, unchanged; the radial output is also unchanged because its definition centers the input. Pairwise output differences are therefore gauge invariant, and the displayed witness $z_q = L(1/4, -1/4, 0, \ldots, 0)$ is centered.

Put

$$Z_q = e^{L/4} + e^{-L/4} + K - 2.$$

The definition of L gives $b = (1-q)e^{-L}$. For $i \in \{0,1\}$, let

$$\eta_i = \frac{bZ_q}{\alpha e^{(z_q)_i}}.$$

Since $Z_q \le Ke^{L/4}$,

$$0 \le \eta_0 \le \frac{K(1-q)}{\alpha} e^{-L}, \qquad 0 \le \eta_1 \le \frac{K(1-q)}{\alpha} e^{-L/2}. \tag{36}$$

For fixed K, both vanish as $q \to 0^+$. Common output centering cancels in the difference for the chosen pair, so

$$\begin{aligned}(\Phi_q(z_q))_0 - (\Phi_q(z_q))_1 &= \frac{L}{2} + \log(1+\eta_0) - \log(1+\eta_1) \\ &= \frac{L}{2} + o(1),\end{aligned} \tag{37}$$

where the absolute remainder is at most $\eta_0 + \eta_1$.

Because $r(z_q) = L/2$, the radial gap for the chosen pair is exactly

$$(R_q(z_q))_0 - (R_q(z_q))_1 = L \tanh(\alpha/2). \tag{38}$$

Subtracting (38) from (37), dividing by L, and using $\alpha \to 1$ yields

$$\frac{[(\Phi_q(z_q))_0 - (\Phi_q(z_q))_1] - [(R_q(z_q))_0 - (R_q(z_q))_1]}{L} \longrightarrow \frac{1}{2} - \tanh\frac{1}{2} = c_\star > 0. \tag{39}$$

The positivity follows from $\tanh x < x$ for $x > 0$. The eventual lower bound and divergence of the centered supremum follow directly from (39).

It remains to compare the decoded distributions at the same witness. Coordinate 0 is the unique maximizer of both outputs. For $K \ge 3$, every zero coordinate has

$$\eta_{\text{zero}} = \frac{bZ_q}{\alpha} \le \frac{K(1-q)}{\alpha} e^{-3L/4} \longrightarrow 0. \tag{40}$$

Hence the exact margin of the top over zero is $L/4 + o(1)$, while (37) gives the exact margin of the top over coordinate 1 as $L/2 + o(1)$. When $K = 2$ there are no zero coordinates, and (37) is the only margin over a wrong state. Every exact margin of the top over a wrong state is therefore eventually at least $L/5$.

Under the radial map, every margin of the top over a wrong state is at least $(L/2)\tanh(\alpha/2) \ge L/5$ eventually. The exact transition probabilities lie strictly between b and $1-q$, so its pairwise logit diameter is below L; the radial pairwise diameter is also below L. Thus the range of $\Phi_q(z_q) - R_q(z_q)$ is below $2L$. Applying Lemma 20 with $m = L/5$ and $R = 2L$ gives

$$D_{\mathrm{KL}}(\operatorname{softmax}\Phi_q(z_q) \,\|\, \operatorname{softmax} R_q(z_q)) \le 4K(K-1)L^2 e^{-L/5} \longrightarrow 0. \tag{41}$$

This proves all claims. □

E.9 Proof of the binary transition geometry

Proof of Proposition 6. Writing $u = e^h$ and $D_q(u) = ((1-q)u + q)(qu + 1 - q)$ gives

$$F_q'(h) = \frac{(1-2q)u}{D_q(u)}, \qquad F_q''(h) = \frac{(1-2q)q(1-q)u(1-u^2)}{D_q(u)^2}.$$

The first formula, equivalently $F_q'(h) = (1-2q)p(1-p)/[p'(1-p')]$ for $p = \sigma(h)$ and $p' = q + (1-2q)p$, is at most $1-2q$, with equality only at $h = 0$; the range $p' \in (q, 1-q)$ proves saturation. The second formula is strictly negative for $h > 0$, so strict concavity on $[0, 2x]$ and $F_q(0) = 0$ give $\Delta_q = 2F_q(x) - F_q(2x) > 0$. For residuals $e_z = F_q(z) - A(z)$, the affine second difference vanishes and hence

$$\Delta_q = |e_0 - 2e_x + e_{2x}| \le 4 \max_z |e_z|,$$

which proves the quantitative obstruction. □

Table 5: **Equally spaced Gaussian illustration at fixed scale.** Internal slopes regress centered separation from exact to radial on L; slopes of the log KL regress $\log D_{\mathrm{KL}}(\mathrm{exact}\|\mathrm{radial})$ on L. These values illustrate the theorem's two directions for three fixed state counts; they are not used in its proof.

K	Internal slope [95% CI]	Final gap/L [95% CI]	Slope of log KL [95% CI]
2	0.5794 [0.5788, 0.5800]	0.5419 [0.5414, 0.5423]	−0.1574 [−0.1661, −0.1454]
4	0.7493 [0.7487, 0.7499]	0.7315 [0.7311, 0.7319]	−0.0265 [−0.0292, −0.0241]
8	0.8688 [0.8681, 0.8696]	0.8728 [0.8724, 0.8733]	−0.0085 [−0.0090, −0.0080]

F Empirical Estimands and Aggregation

This appendix defines the illustration with equally spaced Gaussians at fixed K, experiments with learned models, and estimands for the scalar controls. These experiments illustrate the results and are not used in the proof of Theorem 4. The displayed subset with equally spaced Gaussians was selected after the broader prespecified grid had been evaluated, to illustrate the two directions rather than test a uniform onset at finite scale. On the original Gaussian grid, four of five configurations pass the joint directional criterion; the $K = 8$ skewed configuration has an unresolved slope of the log KL over the prespecified tail. The supplement includes results for the full original grid (`empirical/fixed-k-bridge/results/full/`) and identifies the additional experiments conducted afterward (`empirical/fixed-k-bridge/results/k8-extension/`); the displayed subset does not replace the original outcome.

F.1 Illustration with equally spaced Gaussians at fixed K

F.1.1 Design, coupling, and estimands

For $K \in \{2, 4, 8\}$, the scalar Gaussian means begin as the arithmetic progression $0, 1, \ldots, K-1$, are centered, and are rescaled so that the minimum pairwise distance is one. The observation variance is one, hence

$$d_{\min} = \min_{i \neq j} \frac{(\mu_i - \mu_j)^2}{2} = \frac{1}{2}, \qquad c = 0.45 d_{\min} = 0.225.$$

At each $L \in \{8, 16, 32, 64, 128\}$ we set

$$q = \frac{K-1}{e^L + K - 1}, \qquad H = \left\lceil \frac{L}{0.225} \right\rceil + 1.$$

This implemented horizon differs from the theorem's $\lceil -\log(q)/c \rceil + 1$ by $O_K(1)$ steps, since $L = -\log q + \log(K-1) + \log(1-q)$; it shares the logarithmic scale rather than the exact schedule. Four seed blocks contribute 1,024 stationary trajectories per cell, for 4,096 paired paths. Exact Bayes and the tanh radial filter consume the same observations. Every displayed path and filter state is finite.

The internal quantity is $G_H = \|\mathcal{C}(z_H^E - z_H^R)\|_2$ and the task quantity is $D_{\mathrm{KL}}(p_H^E \| p_H^R)$. The normalized gap G_H/L distinguishes growth proportional to the confidence scale from a merely positive absolute gap. Internal slopes regress mean G_H on L; predictive slopes regress the logarithm of mean predictive KL on L. Both use the predeclared tail window $L \in \{32, 64, 128\}$.

Uncertainty is computed by resampling complete paths. Each of 2,000 bootstrap replicates independently resamples path indices within each scale cell, then fits the statistic across those resampled cell means. Filters remain paired on each path, but independently simulated scales are not paired. Table 5 reports the three displayed configurations.

F.1.2 Controls and implementation details

Hard clipping and the affine update are comparisons rather than assumptions of the theorem. Hard clipping may outperform tanh at finite scales because the theorem claims existence and asymptotic convergence, not optimality of the smooth map. The affine control has lower mean KL than tanh in every displayed tail cell $L \geq 32$. All 4,096 sampled paths in each such cell contain no switch even

by $2H$, so these curves do not test rare costs in the switch tail or identify saturation as necessary. The binary $\lceil 8/q \rceil$ controls address a separate recovery question at the longer endpoint.

The supplementary code includes the data generator, simulation parameters, seed construction, measurements per trajectory, and the 2,000-replicate bootstrap procedure used to produce the curves and table. Numerical checks verify the KL direction and that all simulated trajectories are complete and finite. Table values are rounded to four decimal places; the accompanying data provide full precision.

F.2 Experiments with learned models

The experiment covers six switch probabilities $q \in \{2^{-3}, \ldots, 2^{-8}\}$, two training stages (G1 distilled and G2 trained end to end), five learned arms and an additional analytic tanh reference, three tuning seeds, and ten disjoint confirmation seeds. For each stage and model, the selected hyperparameter setting minimizes mean validation loss over all three tuning seeds, with ties broken deterministically; a setting missing any tuning seed is ineligible. This selection rule was fixed before confirmation. All 180 tuning runs and 600 final evaluation runs completed successfully. Every confirmation seed is included in the reported aggregates. The experimental protocol also specified test excess NLL on the next observation and calibration error, but neither was measured. Table 6 specifies the model architectures and selected optimizer settings; the official Mamba implementation is `mamba-ssm==2.3.2.post1`.

The closure statistic in Table 2 is computed within a common stage, q, seed, and endpoint:

$$C_a = \frac{D_{\mathrm{S6}} - D_a}{D_{\mathrm{S6}} - D_{\tanh}}.$$

Thus S6 is 0 and the analytic tanh arm is 1 before averaging. The table reports the mean and standard error of these ten closure values, one for each seed, not a standard error obtained by treating paths within a seed as independent replicates.

F.3 Predictive KL: orientation and clipping

The experiments with learned models and the scalar controls also use different probability clipping settings. Table 2 uses $D_{\mathrm{KL}}(\text{exact Bayes} \| \text{model})$ after float32 sigmoid conversion and clipping at $\varepsilon_{32} = \texttt{torch.finfo(torch.float32).eps} \approx 1.1920929 \times 10^{-7}$. The scalar controls below use float64 probabilities clipped at 10^{-15}. These settings compute KL after different clipping operations, so the reported values are not directly interchangeable.

The fixed-K theorems, Table 7, and Figure 2 all use the direction $D_{\mathrm{KL}}(\text{exact} \| \text{approximation})$, although the experiments use their separately disclosed probability clipping settings. An additional diagnostic reverses the direction and reports $D_{\mathrm{KL}}(\tanh \| \text{exact})$. Because KL is asymmetric, those values in the reverse orientation are not compared numerically with the theorem or the comparison from exact to control.

That auxiliary diagnostic evaluates the tanh recurrence at $H(q)$ without probability clipping. It computes Bernoulli KL directly in logit space, uses the same ten seeds and 20,000 paths per seed, and is reported separately from the comparison from exact to control.

q	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}
$D_{\mathrm{KL}}(\tanh \| \text{exact})$	1.74×10^{-4}	8.53×10^{-4}	2.27×10^{-3}	4.23×10^{-3}	6.00×10^{-3}	7.63×10^{-3}

The largest relative standard error is 0.40%. The loss in the reverse orientation increases across this finite grid. It is a diagnostic in the finite regime, not an estimate of either proved limit from exact to radial or its asymptotic rate.

If e is the exact predictive logit, m is a model predictive logit, $p = \sigma(e)$, and $r = \sigma(m)$, the corresponding untruncated empirical estimand is

$$D_{\mathrm{true}}(e\|m) = D_{\mathrm{KL}}(\mathrm{Ber}(p)\|\mathrm{Ber}(r)) = p(e-m) + \mathrm{sp}(m) - \mathrm{sp}(e), \qquad \mathrm{sp}(z) = \log(1+e^z).$$

The last expression is a stable evaluation in logit space of the ordinary Bernoulli KL; implementations may evaluate sp by `logaddexp` without clipping either probability.

Table 6: **Model architectures and training settings.** All arms use AdamW, batch size 256, at most 10,000 updates, validation every 250 updates, and patience 2,000; η and λ denote learning rate and weight decay. G1 distills the exact predictive distribution, whereas G2 minimizes negative log likelihood of the next observation. Parameter counts are matched within 5%. The arm with a finite window uses $H(q) = \lceil -\log q \rceil + 1$; the Sharan existence bound $\log 2/H(q)$ ranges from 0.173 to 0.099 over this grid and is not a guarantee for the trained MLP.

Arm	Architecture	Parameters	G1 / G2 (η, λ)
Mamba $\times 2$	two blocks; $d_{\text{model}} = 32$, $d_{\text{state}} = 16$, convolution 4, expansion 2	20,001	$(10^{-3}, 10^{-2})$ / $(10^{-4}, 0)$
Mamba $\times 1$	one block; $d_{\text{model}} = 48$, otherwise as above	19,873	$(10^{-3}, 0)$ / $(3 \times 10^{-4}, 0)$
GRU	one layer, hidden width 80	20,001	$(10^{-3}, 0)$ / $(10^{-3}, 0)$
Scalar S6	one scalar state; two layers of width 139 in the SiLU coefficient network	20,018	$(3 \times 10^{-4}, 0)$ / $(3 \times 10^{-4}, 0)$
Finite window	last $H(q)$ observations plus validity mask; widths 136, 136, 135, 134, 134, 133 for $q = 2^{-3}, \ldots, 2^{-8}$	19,951–19,993	$(10^{-3}, 0)$ / $(3 \times 10^{-4}, 0)$

Table 7: **Controls separate saturation from geometry.** Expected KL at the final step, KL(exact Bayes‖control), at the endpoint matched to the theorem $H(q) = \lceil -\log q \rceil + 1$ and the long endpoint $\lceil 8/q \rceil$. At the long endpoint, the saturating losses remain below 6.4×10^{-3}, while the tested affine and identity losses reach 5.53 and 14.54 nats, respectively (bold).

	Saturating		Nonsaturating	
Endpoint	**tanh**	**clip**	**affine**	**identity**
$q = 2^{-8}$				
$H(q)$	6.0×10^{-3}	1.3×10^{-3}	7.9×10^{-2}	8.4×10^{-2}
$\lceil 8/q \rceil$	6.36×10^{-3}	1.37×10^{-3}	**5.53**	**14.54**
$q = 2^{-3}$				
$H(q)$	1.72×10^{-4}	3.20×10^{-3}	2.44×10^{-1}	9.15×10^{-1}
$\lceil 8/q \rceil$	1.73×10^{-4}	3.20×10^{-3}	4.06×10^{-1}	**10.28**

The experiment with scalar controls evaluates a distinct, clipped quantity. For $c_\varepsilon(u) = \min\{1 - \varepsilon, \max\{\varepsilon, u\}\}$ with $\varepsilon = 10^{-15}$, it records

$$D_{\text{clip},\varepsilon}(e \| m) = D_{\text{KL}}(\text{Ber}(c_\varepsilon(\sigma(e))) \| \text{Ber}(c_\varepsilon(\sigma(m)))) .$$

Accordingly, every scalar value currently printed in Table 7 and plotted in the third panel of Figure 2 is clipped predictive KL with orientation *exact Bayes ‖ control*. Clipping is nearly inactive for the two saturating controls but can substantially cap a nonsaturating control in an extreme tail. The clipped and untruncated estimands must therefore be named separately. The reported control results use clipped KL, not the untruncated quantity.

F.4 Scalar controls and endpoints

All four scalar recurrences consume the same simulated observation paths as the exact filter. After T observations, let h_T be the exact posterior logit and $\widetilde{h}_T^a$ the state of control a. The final logits compared by the scalar evaluator are

$$e_T = F_q(h_T), \qquad m_T^a = M_q^a(\widetilde{h}_T^a), \qquad a \in \{\text{tanh}, \text{clip}, \text{affine}, \text{identity}\}.$$

The two endpoints are the horizon matched to the theorem $H(q) = \lceil -\log q \rceil + 1$ and the long endpoint $\lceil 8/q \rceil$. The experiment in Table 7 and Figure 2 uses $q = 2^{-3}$ and $q = 2^{-8}$, ten public seeds $1080, \ldots, 1089$, 20,000 paths per seed at $H(q)$, and 2,000 paths per seed at the long endpoint. Thus each short cell contains 200,000 paths and each long cell contains 20,000 paths.

For a seed s with N_s paths, the stored replicate is

$$\widehat{D}_s^a = \frac{1}{N_s} \sum_{i=1}^{N_s} D_{\mathrm{clip},10^{-15}}(e_{T,i} \| m_{T,i}^a).$$

With $S = 10$, the manuscript mean and uncertainty are

$$\overline{D}^a = \frac{1}{S} \sum_{s=1}^{S} \widehat{D}_s^a, \qquad \mathrm{SE}(\overline{D}^a) = \sqrt{\frac{1}{S(S-1)} \sum_{s=1}^{S} (\widehat{D}_s^a - \overline{D}^a)^2}.$$

The random stream for one q/endpoint/seed cell is derived from the displayed seed, the exponent of q, and the horizon; all four controls reuse that cell. The manuscript aggregates seed means rather than treating paths within a seed as independent replicates. The largest relative standard error among the sixteen plotted quantities is at most 3.1%.

Forced-switch recovery diagnostic. To test the boundary excluded by Theorem 4, we also force exactly one binary-state switch after $H(q)$ observations and then hold the new state fixed for another $H(q)$ observations. Over 20,000 paths at $q \in \{2^{-3}, 2^{-5}, 2^{-8}\}$, the exact and radial filters end with mean mass on the new state $(0.881, 0.929, 0.958)$ and $(0.879, 0.928, 0.963)$, respectively; their terminal KL is 7.13×10^{-5}, 1.05×10^{-3}, and 6.59×10^{-3}. This deliberately conditioned experiment diagnoses recovery after a switch; it is not sampled from the stationary path law and is not used in the theorem.

The right pair in Figure 2 separates measurement from illustration. The third panel reports the mean losses of the scalar controls. The fourth panel is not a sampled trajectory: it evaluates the logistic decoder on a fixed grid, with schematic guide logits $v = 3$ and $u = 5.25$.

F.5 Errors relative to the latent state

Let $Z_{T+1} \in \{0, 1\}$ be the latent state after the terminal transition, $s_Z = 2Z_{T+1} - 1$, e_T the exact predictive logit, and m_T^{aff} the affine predictive logit. The event that the affine prediction disagrees with the latent state is

$$W_{\mathrm{truth}} = \mathbf{1}\{s_Z m_T^{\mathrm{aff}} \le 0\}.$$

The associated confidence event used by the formal diagnostic for the lower bound is

$$W_{\mathrm{joint}} = W_{\mathrm{truth}}\, \mathbf{1}\{\sigma(s_Z e_T) \ge 15/16\}.$$

Both indicators are defined relative to the true latent state. Disagreement with the exact filter is a different quantity and was not measured in this experiment.

The experiment uses $q = 2^{-5}, 2^{-8}, 2^{-10}, 2^{-12}$, the same ten public seeds, and 1,000 paths per seed. Means across seeds $\pm$ standard errors for W_{truth} are respectively 0.1982 ± 0.0064, 0.1821 ± 0.0042, 0.1829 ± 0.0031, and 0.1868 ± 0.0036. The corresponding W_{joint} values, qualified by confidence, are 0.0794 ± 0.0027, 0.1649 ± 0.0046, 0.1781 ± 0.0028, and 0.1856 ± 0.0036. The reported range describes errors relative to the latent state, rather than disagreement with the exact filter.

F.6 Scope of the empirical conclusions

The experiments support two deliberately limited conclusions. First, at the long endpoint the saturating controls remain stable while the two nonsaturating controls degrade under the disclosed clipped metric. Second, matching the short theorem horizon does not by itself establish stability at long horizons for a learned architecture. Neither observation proves the open lower bound for the terminal joint event qualified by confidence, identifies tanh as the unique stable recurrence, or establishes a positive result on transfer between architectures.

G Extended Background

This appendix places the result in the literature on recurrent models, approximate filtering, and sequential prediction.

G.1 THE FAMILY WITH A DETERMINISTIC STATE

The recurrence studied here is deliberately the smallest object that exhibits the property in question: a state of fixed size updated deterministically from the previous state and the current observation. For recurrent implementations, this property permits inference in linear time and memory independent of sequence length. Configurations with long convolutions or attention hybrids need not share that memory bound.

The lineage begins with the question of how a state of fixed size should summarize a growing history at all. Optimal polynomial projection gives one answer with an explicit approximation guarantee (Gu et al., 2020), and structured state spaces turn that answer into a trainable layer (Gu et al., 2022), later simplified (Smith et al., 2023) and given variants in continuous time (Hasani et al., 2023). Initialization of these models remains an active question in its own right (Lienen et al., 2025). A parallel line showed that carefully parameterized linear recurrences recover much of the benefit without the state space machinery (Orvieto et al., 2023), and designs with long convolutions and hybrid designs traded recurrence for structured mixing (Poli et al., 2023; Fu et al., 2023). Gated linear attention and its delta rule successors reached the same destination from the attention side (Yang et al., 2024a;b), with sparsity now used to enlarge the state without paying for it densely (Cabannes et al., 2026), while recurrent architectures competitive at language scale arrived independently (Peng et al., 2023; Beck et al., 2024). The selective mechanism that makes the transition depend on the input (Gu & Dao, 2024) and the duality that exposes such models as a restricted form of attention (Dao & Gu, 2024) complete the picture.

These models motivate asking what predictive cost follows from a prescribed update geometry. Shared motivation does not extend our existence theorem to every member of this family: the theorem concerns the explicit radial filter, not its implementation by a named neural architecture.

G.2 WHAT IS ALREADY KNOWN ABOUT THE GAP, AND WHAT IS NOT

The representational side of the question is comparatively well mapped. Formal placements bound what specified architectures can express (Sarrof et al., 2024), while empirical studies on formal languages probe generalization outside the training distribution (Delétang et al., 2023). Arguments from circuit complexity place models of fixed depth inside classes that cannot perform certain sequential computations (Merrill et al., 2024), and analyses with limited precision show how much of the apparent capacity survives finite arithmetic (Li & Cotterell, 2025). On the constructive side, changing the spectrum or the sparsity pattern of the transition recovers ability to track the state that diagonal models lack (Grazzi et al., 2024; Terzić et al., 2025; Siems et al., 2025), and the frontier between recall and throughput has been characterized directly (Arora et al., 2024). Benchmarks now target exact state tracking as a primary object (Walker & Lyons, 2026).

What this literature does not settle is the step our paper takes. A proved inability to represent an update is not by itself a proved cost under the data distribution, because the distribution need not visit the region where the representation fails. The closest existing work asks about generalization of selective state space models on filtering tasks (Tang et al., 2026), which shares our setting but not our question: we fix the recurrence and ask what its representational shortfall costs in expected predictive divergence.

G.3 FILTERING, STABILITY, AND THEIR RELATION TO THIS RESULT

Hidden Markov filtering has a long history (Rabiner, 1989). Exact filtering of a K-state HMM admits a deterministic continuous belief state with $K - 1$ coordinates. Finite dimension is not finite cardinality or bounded precision; our restriction concerns update geometry, not the existence of a sufficient state of finite dimension. The stability literature asks a superficially similar question to ours and a materially different one: whether a filter initialized incorrectly but running the correct kernels forgets its error (Ocone & Pardoux, 1996), surveyed by Chigansky et al. (2011), with quantitative forms via contraction coefficients (McDonald & Yüksel, 2020). Robustness results perturb the kernels instead, controlling policy costs (Kara & Yüksel, 2020; Kara et al., 2022) or error in the filter kernel and policy performance (Demirci et al., 2025). Policies with a finite window can become nearly optimal as window length grows (Kara & Yüksel, 2022). Our state dimension is fixed, but R_q changes numerically with q, and its map discrepancy in the worst case grows rather than vanishing as a perturbation parameter. How fast a correct filter forgets is itself quantified (Le Gland & Oudjane, 2004), and estimating the mixing time is a developed subject (Wolfer & Kontorovich, 2019).

Learned and structured filters form a third strand. Probabilistic state space models have been fused with the selective architecture (Becker et al., 2024), particle filters made differentiable (Corenflos et al., 2021), belief embeddings learned with consistency guarantees (Solinas et al., 2025), and Updates in the style of the Kalman filter recast as attention or as regression at test time (Shaj et al., 2026; Peng et al., 2025), including adaptive variants (Mehrfard et al., 2026). Learning from ordinary HMM observation sequences can be cryptographically hard, whereas query access to conditional probabilities permits efficient learning; guarantees from conditional samples also depend on a fidelity parameter (Kakade et al., 2023). Reconstruction from finite messages on trees exhibits phase transitions between memory and accuracy (Jain et al., 2019), and a recent synthesis compares state space models and HMMs (Ghojogh et al., 2026).

G.4 Prediction under a wrong model

Our theorem measures the categorical KL between filtered posteriors after the observation, not the KL between densities of the next observation, and binary empirical endpoints additionally apply transition mixing. These losses score decoded distributions rather than internal parameters. Prediction with short memory supplies a conceptual benchmark on the next observation (Sharan et al., 2018); sequential prediction under log loss with a misspecified model class is studied directly (Feder & Polyanskiy, 2021); and streaming projection onto a mixture over paths with a fixed budget is another deterministic recurrence with a different maintained object and truncation (Duran-Martin, 2026). Read as detection delay, our controls at long horizons belong to quickest change detection, from the original cumulative sum scheme (Page, 1954) through its optimality theory (Lorden, 1971; Moustakides, 1986; Pollak, 1985) to learned detectors (Gong et al., 2022). Behavior in context on Markov sources is a further contact point (Bondaschi et al., 2025; Zhang & Xing, 2026).

G.5 Comparison by mathematical object

Nearby literatures can sound interchangeable when summarized as "approximating a sequence model," but they differ in what may change with the accuracy target and in where error is measured. Table 8 records the distinction needed here.

Table 8: **Nearby questions separated by object and limit.** The comparison is conceptual; it does not claim that the cited settings are special cases of ours.

Line of work	Object being compared	Accuracy resource or limit	Relation to this paper
Architecture expressivity (Merrill et al., 2024; Sarrof et al., 2024; Cohen-Karlik et al., 2026)	Function or language represented by an architecture class	Depth, state structure, or input length varies with the theorem	Establishes representational separations, but generally does not integrate their task cost under a data law.
Prediction with short memory (Sharan et al., 2018)	Predictor with full history versus a finite observation window	Window length grows as the target error shrinks	Supplies a positive approximation baseline; our state dimension and prescribed family remain fixed, but the numerical map depends on q.
Filter stability and robustness (Ocone & Pardoux, 1996; McDonald & Yüksel, 2020; Le Gland & Oudjane, 2004)	Two filters with different initial beliefs or nearby kernels	Influence of the initial condition is forgotten or kernel perturbation is taken small	Controls belief error when the source of mismatch disappears; our two transition geometries do not converge internally.
Compression that is aware of the task (Dubois et al., 2021; 2020; Hafez-Kolahi et al., 2021)	Original representation versus a compressed representation	Compression is optimized relative to a downstream task	Shares the lesson that discarded information can be irrelevant to the task, but not the recurrent Bayesian construction under the path law.
This paper	Exact categorical Bayes mixing versus one radial update with a fixed state	Fixed finite K; switch rarity tends to zero at $H(q) = \Theta(\log(1/q))$	Proves diverging centered map separation and vanishing decoded KL, then proves stationary expected predictive closure.

The comparison also limits the architectural interpretation: a selective SSM, a gated recurrent network, and the radial filter all keep a fixed state, but the theorem concerns the explicit radial map, not a class containing every such recurrence. Exact Bayes and the radial witness are mathematical filters, so their separation predates training or parameterization. The result sits between architecture expressivity and statistical decision theory: an internal separation can fail to induce predictive loss under a specified sequential law, leaving the converse criterion and learned realization open.