A COMPARATIVE ANALYSIS OF ATTENTION VERSUS STATE-SPACE MODELS FOR IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformers and state-space models (SSMs) are two prominent sequential learning architectures, yet their comparison remains largely empirical and existing theoretical analyses are typically task-specific or architecturally restricted. In this paper, we develop *belief geometry*, a unified analytical framework for comparing the representational capabilities of broad classes of attention and SSMs. Starting from a generalized formulation of in-context linear regression and using cumulative Bayes regret as our measure, we abstract three capabilities required by many sequential learning problems in our belief geometry: evidence assembly, belief maintenance, and addressing. We then study three cases of our formulation that isolate these capabilities and yield sharp architectural lessons: For belief maintenance, SSMs attain the optimal regret over stationary aggregation kernels; for positional assembly, SSMs have a memory advantage; and for content addressing, softmax attention has an exponential width advantage over sigmoid-selective SSMs. Experiments with LLaMA-type Transformers and Mamba-2 show that these architectural insights extend beyond our analytically tractable classes and linear-regression testbed.

1 INTRODUCTION

Transformers (Vaswani et al., 2017) and state-space models (SSMs) (Gu & Dao, 2024; Beck et al., 2024; Yang et al., 2024) have emerged as two competitive architecture families for sequential learning. They organize history in fundamentally different ways: Transformers retain past tokens within a context window for direct access, whereas SSMs compress the past into a recurrent state. A principled understanding of when either strategy is advantageous, and which architectural mechanisms give rise to those advantages, has therefore become essential for designing sequential learning architectures.

Comparisons between Transformers and SSMs, however, remain largely empirical (e.g., Arora et al., 2023; Park et al., 2024). Existing theoretical work typically studies specific tasks (e.g., Sanford et al., 2023; 2024b; Jelassi et al., 2024; Wen et al., 2024), analyzes restricted forms of these architectural classes (e.g., Cole et al., 2025; Zhang et al., 2023; Tian et al., 2026), or unifies their mathematical descriptions without explaining when their architectural differences matter (e.g., Sieber et al., 2024). We therefore lack a general framework for identifying which architectural capabilities a sequential learning task demands, how different architectures supply them, and at what resource cost.

Contributions. In this paper, we generalize in-context linear regression (ICLR) through routing and evaluate architectures by their cumulative Bayes regret rather than by their loss at a fixed prediction step. This extension reveals three capabilities required by many sequential learning problems: (i) *assembling evidence* from inputs, (ii) *maintaining a belief* by aggregating that evidence, and (iii) *addressing* the relevant part of history to which that belief is applied. We formalize these capabilities in our *belief geometry*, which enables a clean, unified representational analysis of architecturally rich and practically relevant classes of attention (Transformers' backbone) and SSMs (see Remark 1).

Our analysis proceeds through controlled instantiations of the generalized ICLR setup: in each case, a different capability becomes the bottleneck, and we analyze it through its corresponding element in belief geometry. Starting from the simplest versions of each family, we use linear regret lower bounds to identify the failures and sublinear regret constructions to show which architectural components overcome them. The resulting sharp comparisons of achievable regret and of the depth, width, and memory required by each family explain *where* and *why* one family prevails over the other and yield the following architectural lessons:

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.

- **Belief maintenance: SSMs attain the stationary optimum.** Case 1 isolates belief maintenance, captured in belief geometry by the *aggregation kernel*. Our key observation is that one-layer linear attention and one-layer fixed-transition SSMs can at best realize stationary kernels. Within this restriction, their best achievable regrets differ: the SSM's exponential kernel attains the stationary optimum, whereas attention's uniform kernel cannot. This separation predicts which kernel practical models in each family favor—an inductive-bias difference that we validate empirically.
- **Positional assembly: SSMs have a memory advantage.** Case 2.a introduces positional routing and makes evidence assembly the bottleneck, captured in belief geometry by *belief alignment*. The Case 1 architectures incur linear regret at any width or context length because retaining history cannot substitute for assembly. Attention therefore needs depth and uses its context length for both reach and belief maintenance, whereas the SSM separates them through convolution and recurrence, attaining the same regret rate with *strictly less memory*.
- **Content addressing: attention has an exponential width advantage.** Case 2.b introduces content routing and makes addressing the bottleneck, quantified in belief geometry by *addressing capacity*. The Case 2.a mechanisms cannot sharply address the required historical input: at any fixed depth, linear attention and affine-gated SSMs incur linear regret regardless of width or context length. The families escape this floor differently: attention uses softmax to select among retained tokens, whereas the SSM uses sigmoid-selective transitions to preserve candidates separately in its recurrent state and requires *exponentially larger width* for the same regret rate.

Finally, experiments with LLaMA-type Transformers and Mamba-2 show that these lessons extend beyond our analytically tractable classes and linear-regression testbed. Appendix D.3 additionally provides direct experimental validation of our theoretical results, which are summarized in Table 1.

Related work. Within the theory of in-context learning, work closest to our setup characterizes which regression algorithms attention and recurrent models can realize in ICLR, typically studying one family at a time and evaluating loss at a fixed prediction step (von Oswald et al., 2023; Mahankali et al., 2023; Bai et al., 2023; Tian et al., 2026; Cole et al., 2025). Our belief-maintenance analysis (Case 1, §3) instead compares *both families* under *cumulative Bayes regret*. For positional assembly (Case 2.a, §4.1), prior work studies attention depth (Sanford et al., 2024a; Ekbote et al., 2025) and convolution in Mamba (Bondaschi et al., 2026); we prove *class-level regret floors* and an *SSM memory advantage* at the same sublinear regret rate. The literature closest to content addressing (Case 2.b, §4.2) establishes attention–recurrence separations for associative recall, induction heads, and copying under computational, memory, or statistical criteria (Sanford et al., 2023; 2024b; Jelassi et al., 2024; Wen et al., 2024; Mousavi-Hosseini et al., 2025). These tasks are largely standalone and discrete, and their lower bounds measure state size in bits under a finite-precision assumption. Our addressing problem is instead embedded in noisy Bayesian inference alongside evidence assembly and belief maintenance; its lower bound concerns *real-valued state dimension* without a precision restriction. At a broader theoretical level, prior work unifies the mathematical representation of attention and SSMs (Sieber et al., 2024; Wang et al., 2025). Our belief geometry instead *unifies their analysis*, translating architectural differences into task-dependent regret and resource comparisons. Appendix A compares our results with these works in greater detail and surveys the broader literature.

2 Setup

Data-generating process. We generalize Bayesian in-context linear regression (ICLR) by introducing routing into the setup studied in prior work (Garg et al., 2023; Akyürek et al., 2023; von Oswald et al., 2023; Zhang et al., 2023). A partial injective *routing map* $\rho : \mathbb{N} \rightharpoonup \mathbb{N}$ specifies which regressor generates the response $y_s \in \mathbb{R}$ for each time step s in the domain of ρ, subject to the causal constraint $\rho(s) \leq s$. Whenever $\rho(s)$ is defined, step s is called *matched*, and at that step the data stream obeys

$$y_s = x_{\rho(s)}^\top \theta + \epsilon_s, \qquad \theta \sim \mathcal{N}(0, \Sigma_\theta), \quad x_s \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_x), \quad \epsilon_s \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

where $\theta, x_s \in \mathbb{R}^p$ for a fixed $p \in \mathbb{N}$, $\Sigma_\theta, \Sigma_x \succ 0$, $\sigma^2 > 0$, and $\theta, (x_s)_s, (\epsilon_s)_s$ are mutually independent. All remaining steps are unmatched, with $y_s = 0$, and therefore carry no information about θ. Throughout, expressions involving $\rho(s)$ are understood only for matched s. At step t, the task is to predict y_{t+1} from the history $z_{1:t}$, where $z_s = (x_{s+1}, y_s, x_s) \in \mathbb{R}^{2p+1}$. So at every step, the history contains the regressor $x_{\rho(t+1)}$ that generates y_{t+1}, but not the response itself. This lets us evaluate cumulative Bayes regret rather than only the loss at the end of the context.

Bayes regret. Let $\mathcal{H}_t := L^2(\sigma(z_{1:t}))$, with its usual inner product $\langle \cdot, \cdot \rangle_t$ and norm $\| \cdot \|_t$. A fixed predictor g maps each history $z_{1:t}$, of any length, to $g_t := g(z_{1:t}) \in \mathcal{H}_t$ and incurs the squared loss $\ell_t(g) := \mathbb{E}[(g_t - y_{t+1})^2]$. The Bayes predictor g^{Bayes} minimizes this loss with $g_t^{\mathrm{Bayes}} := \mathbb{E}[y_{t+1} \mid z_{1:t}]$. Taking it as the comparator, the cumulative regret of a predictor g, and of a class $\mathcal{F}$, over horizon T is

$$\mathrm{Reg}_T(g) := \sum_{t=1}^{T} \left[\ell_t(g) - \ell_t(g^{\mathrm{Bayes}})\right] = \sum_{t=1}^{T} \|g_t - g_t^{\mathrm{Bayes}}\|_t^2, \qquad \mathrm{Reg}_T(\mathcal{F}) := \inf_{g \in \mathcal{F}} \mathrm{Reg}_T(g).$$

Unlike fixed-context ICLR formulations that evaluate a single loss $\ell_T(g)$ after observing the entire context (Bai et al., 2023; Mahankali et al., 2023; Ahn et al., 2023; Wu et al., 2024), our online formulation evaluates the *cumulative excess loss* as the context grows. This captures an essential aspect of sequential learning that terminal-only evaluation misses: performance throughout the entire trajectory, not merely at its endpoint.

Belief geometry. Our belief geometry is motivated by the following structure of the Bayes predictor under routing: it factorizes as $g_t^{\mathrm{Bayes}} = x_{\rho(t+1)}^\top \hat{\theta}_t^{\mathrm{Bayes}}$, where $\hat{\theta}_t^{\mathrm{Bayes}} := \mathbb{E}[\theta \mid z_{1:t}]$ is the Bayes belief, which in turn can be written as an aggregation of evidence atoms through a kernel:

$$\hat{\theta}_t^{\mathrm{Bayes}} = \sum_{s \le t} K_{t,s}^{\mathrm{Bayes}} x_{\rho(s)} y_s, \qquad K_{t,s}^{\mathrm{Bayes}} := \left(\sigma^2 \Sigma_\theta^{-1} + \sum_{r \le t} x_{\rho(r)} x_{\rho(r)}^\top\right)^{-1}. \tag{1}$$

Together, the factorization and kernel form reveal three operations required to solve many sequential-learning problems: *evidence assembly*, *belief maintenance*, and *addressing*. The Bayes predictor performs all three: it assembles evidence atoms $x_{\rho(s)} y_s$, maintains a belief by weighting these atoms with the aggregation kernel $K_{t,s}^{\mathrm{Bayes}}$, and addresses the relevant part of history, $x_{\rho(t+1)}$, to which that belief is applied. Appendix B shows how these operations extend to broader sequential learning problems, including conjugate and nonconjugate setups, and explains how our regret-analysis framework carries over.

To relate an arbitrary predictor g to the Bayes structure and measure how well it performs these operations, let $\mathcal{G}_t^-$ denote the history with $x_{\rho(t+1)}$ held out and define the *belief subspace*

$$\mathcal{H}_t^\rho := \left\{ x_{\rho(t+1)}^\top a : a \in L^2(\mathcal{G}_t^-; \mathbb{R}^p) \right\} \subseteq \mathcal{H}_t.$$

Projecting g_t onto this subspace gives

$$g_t = x_{\rho(t+1)}^\top \hat{\theta}_t(g) + g_t^\perp,$$

which separates g's maintained belief $\hat{\theta}_t(g)$ from the orthogonal component $g_t^\perp \perp \mathcal{H}_t^\rho$, which cannot arise from applying a belief to the addressed regressor (Figure 1).

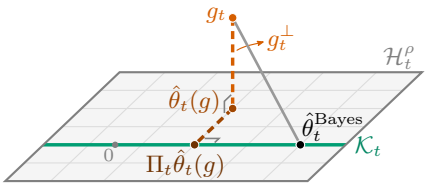


Figure 1: Belief geometry in $\mathcal{H}_t$.

Within $\mathcal{H}_t^\rho$, the Bayes kernel form motivates a further refinement: the subspace of beliefs expressible as weighted sums of evidence atoms. Let $\mathcal{X}_t^- \subseteq \mathcal{G}_t^-$ be the σ-field generated by the regressors in $\mathcal{G}_t^-$, and define the *kernel-form subspace*

$$\mathcal{K}_t := \overline{\left\{ \textstyle\sum_{s \le t} K_{t,s} x_{\rho(s)} y_s : K_{t,s} \in L^\infty(\mathcal{X}_t^-; \mathbb{R}^{p \times p}) \right\}} \subseteq L^2(\mathcal{G}_t^-; \mathbb{R}^p).$$

Let $\Pi_t : L^2(\mathcal{G}_t^-; \mathbb{R}^p) \to \mathcal{K}_t$ denote orthogonal projection under the inner product $\langle a, b \rangle_{\Sigma_x} := \mathbb{E}[a^\top \Sigma_x b]$ and the induced norm $\|a\|_{\Sigma_x}^2 := \mathbb{E}[a^\top \Sigma_x a]$. Since Equation (1) shows that $\hat{\theta}_t^{\mathrm{Bayes}} \in \mathcal{K}_t$, the Pythagorean decomposition visualized in Figure 1 gives

$$\|g_t - g_t^{\mathrm{Bayes}}\|_t^2 = \underbrace{\|g_t^\perp\|_t^2}_{\text{outside belief form}} + \underbrace{\|\hat{\theta}_t(g) - \Pi_t \hat{\theta}_t(g)\|_{\Sigma_x}^2}_{\text{outside kernel form}} + \underbrace{\|\Pi_t \hat{\theta}_t(g) - \hat{\theta}_t^{\mathrm{Bayes}}\|_{\Sigma_x}^2}_{\text{kernel mismatch}}. \tag{2}$$

When an architecture applies a kernel-form belief to the addressed regressor, the first two terms vanish and only kernel mismatch remains. Case 1 (§3) begins directly in this regime: trivial routing isolates belief maintenance, and our projection arguments reduce the analysis to the optimal aggregation kernels each architecture can realize. Case 2 (§4) then introduces nontrivial routing, bringing the first two terms into play and motivating new elements of belief geometry that capture failures of evidence assembly and addressing. Before turning to these cases, we define the attention and SSM classes.

2.1 ARCHITECTURE CLASSES AND THEIR COMPONENTS

We study rich classes of attention and SSMs whose components either appear in practice or serve as minimal tractable surrogates (see Appendix C for practical correspondence). Throughout, D, H, d, and d_e denote depth, head count, per-head width, and residual-stream width, respectively; d is attention's query/key/value width and an SSM head's recurrent-state width. Layer and head indices are $r = 1, \dots, D$ and $b = 1, \dots, H$, respectively. Because d_e primarily follows the input-token dimension and is common to both families, we suppress it from the class notation. Lower bounds are uniform over it, while upper bound constructions assume sufficient width. The displayed arguments in class notations fix architectural hyperparameters, and each class ranges over all remaining parameters, making our analysis representational. All projections are affine, with biases suppressed from notation. Both families initially embed z_s as $e_s^{(0)} = W_{\mathrm{emb}} z_s \in \mathbb{R}^{d_e}$ and finally read out with $W_{\mathrm{pred}} \in \mathbb{R}^{1 \times d_e}$.

Attention ($\mathcal{F}_{\mathrm{att}}(D, d, H, L)$)**.** At time t, attention accesses the causal window $J_t := \{\max(1, t - L + 1), \dots, t\}$. Its parameters at layer r and head b are $W_Q^{(r,b)}, W_K^{(r,b)}, W_V^{(r,b)} \in \mathbb{R}^{d \times d_e}$, $W_O^{(r,b)} \in \mathbb{R}^{d_e \times d}$, and $p_\ell^{(r,b)} \in \mathbb{R}$ for $0 \le \ell < L$. It computes

$$s_{t,s}^{(r,b)} = \left\langle W_Q^{(r,b)} e_t^{(r-1)}, W_K^{(r,b)} e_s^{(r-1)} \right\rangle + p_{t-s}^{(r,b)}, \quad s \in J_t, \qquad a_t^{(r,b)} \in \left\{ s_t^{(r,b)}, \mathrm{softmax}\big(s_t^{(r,b)}\big) \right\},$$
$$o_t^{(r,b)} = \sum_{s \in J_t} a_{t,s}^{(r,b)} W_V^{(r,b)} e_s^{(r-1)}, \qquad e_t^{(r)} = e_t^{(r-1)} + \sum_b W_O^{(r,b)} o_t^{(r,b)}, \qquad g_t = W_{\mathrm{pred}} e_t^{(D)}.$$

Choosing $a_t^{(r,b)} = s_t^{(r,b)}$ gives linear attention, whereas choosing $a_t^{(r,b)} = \mathrm{softmax}(s_t^{(r,b)})$ gives softmax attention; $\mathcal{F}_{\mathrm{att}}$ uses the former throughout, while $\mathcal{F}_{\mathrm{att}}^{\mathrm{sm}}$ permits either. Depth composes successive attention operations, L bounds the accessible history, and p_{t-s} is a tractable additive analogue of the relative-lag dependence induced by RoPE in LLaMA-type Transformers.

State-space models ($\mathcal{F}_{\mathrm{SSM}}(D, d, H, d_{\mathrm{conv}})$)**.** At time t, each SSM layer causally convolves its input, uses the result to write to a recurrent state, and reads that state back to the residual stream. At layer r, head b has recurrent state $h_t^{(r,b)} \in \mathbb{R}^d$, convolution coefficients $C_j^{(r,b)} \in \mathbb{R}^{d_e}$ ($0 \le j < d_{\mathrm{conv}}$), projections $W_K^{(r,b)}, W_V^{(r,b)}, W_Q^{(r,b)} \in \mathbb{R}^{d \times d_e}$ and $W_O^{(r,b)} \in \mathbb{R}^{d_e \times d}$, and transition $\Phi_t^{(r,b)} \in \mathbb{R}^{d \times d}$. Initial states and token padding are zero. With $\odot$ denoting coordinatewise multiplication, it computes:

$$\widetilde{e}_t^{(r,b)} = \sum_{j=0}^{d_{\mathrm{conv}}-1} C_j^{(r,b)} \odot e_{t-j}^{(r-1)}, \; u_t^{(r,b)} = (W_K^{(r,b)} \widetilde{e}_t^{(r,b)}) \odot (W_V^{(r,b)} \widetilde{e}_t^{(r,b)}), \; h_t^{(r,b)} = \Phi_t^{(r,b)} h_{t-1}^{(r,b)} + u_t^{(r,b)},$$
$$o_t^{(r,b)} = (W_Q^{(r,b)} \widetilde{e}_t^{(r,b)}) \odot h_t^{(r,b)}, \qquad e_t^{(r)} = e_t^{(r-1)} + \sum_b W_O^{(r,b)} o_t^{(r,b)}, \qquad g_t = W_{\mathrm{pred}} e_t^{(D)}.$$

This coordinatewise formulation includes outer-product writes and matrix–vector reads without extra recurrent state. The transition is fixed when $\Phi_t = \Phi$ and selective when it depends on $\widetilde{e}_t$. We allow

$$\Phi_t \in \Big\{ \underbrace{\Phi}_{\text{fixed}}, \; \underbrace{\alpha_t I}_{\text{scalar gate}}, \; \underbrace{\mathrm{diag}(\boldsymbol{\alpha}_t)}_{\text{diagonal gate}}, \; \underbrace{I - \beta_t k_t k_t^\top}_{\text{delta rule}}, \; \underbrace{\mathrm{diag}(\boldsymbol{\alpha}_t) + r_t^{\mathrm{out}} (r_t^{\mathrm{in}})^\top}_{\text{generalized delta}} \Big\}.$$

Here Φ is learned but input-independent. The scalar gates α_t, β_t and the coordinates of $\boldsymbol{\alpha}_t$ are sigmoids of affine functions of $\widetilde{e}_t$, while $k_t, r_t^{\mathrm{out}}, r_t^{\mathrm{in}}$ are affine in $\widetilde{e}_t$. $\mathcal{F}_{\mathrm{SSM}}$ uses fixed transitions throughout; $\mathcal{F}_{\mathrm{SSM}}^{\mathrm{sel}}$ permits any displayed transition in its first layer and uses fixed transitions thereafter; and $\mathcal{F}_{\mathrm{SSM}}^{\mathrm{aff}}$ permits any displayed transition in every layer with each sigmoid replaced by the identity.

Practical counterparts. For attention, our main-text experiments use LLaMA-type Transformers (Touvron et al., 2023) with causal softmax attention, RoPE, and SwiGLU MLPs. For SSMs, we use Mamba-2 (Dao & Gu, 2024) with scalar-gated selective transitions and additional write and read nonlinearities. Both include RMSNorm and residual connections.

Remark 1 (Architectural scope)**.** *One-layer linear attention $\mathcal{F}_{\mathrm{att}}(1, d, H, L)$ is standard in theoretical ICL (e.g., von Oswald et al., 2023), with one-layer fixed-transition SSMs $\mathcal{F}_{\mathrm{SSM}}(1, d, H, 1)$ providing its recurrent counterpart (e.g., Katharopoulos et al., 2020). Case 1 begins with these models, and as new demands arise, later cases expand the comparison to our substantially richer full classes: both families vary depth and width; attention also allows variable context length, positional bias, and linear or softmax attention, while SSMs allow convolution and selective transitions.*

3 CASE 1: BELIEF MAINTENANCE

Case 1 recovers the canonical Bayesian ICLR setup with the routing map $\rho(s) = s$. Under this route, each past token z_s contains both factors of its evidence atom $x_s y_s$, and the current token z_t contains the next target's regressor x_{t+1}. Evidence assembly and addressing therefore pose no additional demands, leaving belief maintenance as the focus of Case 1. Accordingly, the *aggregation kernel* introduced by our belief geometry becomes the central tool: we use the kernel-mismatch term in (2) to compare the two families by how well their kernels approximate the Bayes kernel.

Stationary kernels. Applying the kernel-form projection in Figure 1 to the simplest classes of both families—one-layer linear attention $\mathcal{F}_{\mathrm{att}}(1, d, H, L)$ and one-layer fixed-transition SSMs $\mathcal{F}_{\mathrm{SSM}}(1, d, H, 1)$—reveals our key structural observation: they can at best realize *stationary aggregation kernels*, $K_{t,s} = \kappa_{t-s}$, whose weights depend only on lag and not on the current time or realized history (Appendix E.1). Larger width or context length can enrich the stationary profiles they realize, but cannot remove this restriction. We therefore first optimize over all stationary kernels to obtain an architecture-independent benchmark before returning to comparing the two families.

Because our performance measure is cumulative regret, a stationary kernel's inability to adapt to time creates a *bias–variance trade-off*. Spreading weight across larger lags aggregates more evidence atoms and lowers variance but increases cold-start bias before those lags become available; concentrating weight on small lags reduces cold-start bias but aggregates fewer evidence atoms and raises variance. To quantify this trade-off, write $\Sigma_x^{1/2}\Sigma_\theta\Sigma_x^{1/2} = U\,\mathrm{diag}(\lambda_1, \ldots, \lambda_p)U^\top$ for the eigendecomposition of the prior covariance in the Σ_x-geometry. For each eigendirection, set $v_i := \lambda_i + \sum_{j=1}^p \lambda_j + \sigma^2$, the corresponding evidence-innovation variance, and summarize the resulting bias–variance scales by $\zeta := \sum_{i=1}^p \sqrt{\lambda_i v_i}$ and $\bar{\zeta} := \sqrt{(\sum_{i=1}^p \lambda_i)(\sum_{i=1}^p v_i)}$. Cauchy–Schwarz gives $\bar{\zeta} \geq \zeta$. Optimizing the bias–variance trade-off over all stationary kernels yields the following universal regret floor.

Theorem 1 (Stationary floor)**.** *Every g whose kernel is stationary obeys* $\mathrm{Reg}_T(g) \geq \zeta\sqrt{T} + o(\sqrt{T})$.

Therefore, neither class can beat this floor by increasing width or context length; these resources only affect how closely each class approaches it. Rather surprisingly, the proposition below shows that the SSM attains the floor in Thm. 1, whereas attention necessarily pays a larger leading constant.

Proposition 1 (Stationary pair)**.** *When $Hd < p$, both classes incur $\Omega(T)$ regret (uniformly in L for attention). When $Hd \geq p$, their leading-order optimal kernels are:*

(a) $\mathcal{F}_{\mathrm{att}}(1, d, H, L)$*: the* uniform kernel $\kappa_u = \frac{1}{L}\Sigma_x^{-1} + O(\frac{1}{L^2} + \frac{1}{T})$ *for* $0 \leq u < L$ *and* 0 *otherwise.*
(b) $\mathcal{F}_{\mathrm{SSM}}(1, d, H, 1)$*: the* exponential kernel $\kappa_u = \Sigma_x^{-1/2} U\,\mathrm{diag}((1-\alpha_i^\star)(\alpha_i^\star)^u)_{i=1}^p U^\top \Sigma_x^{-1/2}$ *for* $u \geq 0$*, where* $\alpha_i^\star = (2\sqrt{v_i T/\lambda_i} - 1)/(2\sqrt{v_i T/\lambda_i} + 1)$.

For fixed L, $\mathrm{Reg}_T(\mathcal{F}_{\mathrm{att}}(1, d, H, L)) = \Omega(T)$. *The optimal* $L^\star(T) := \sqrt{3T(\sum_i v_i)/(\sum_i \lambda_i)}$ *gives*

$$\mathrm{Reg}_T\big(\mathcal{F}_{\mathrm{att}}(1, d, H, L^\star(T))\big) = \tfrac{2}{\sqrt{3}}\bar{\zeta}\sqrt{T} + o(\sqrt{T}), \quad \mathrm{Reg}_T\big(\mathcal{F}_{\mathrm{SSM}}(1, d, H, 1)\big) = \zeta\sqrt{T} + o(\sqrt{T}).$$

Architectural lessons. The optimal kernels reveal how the two families maintain beliefs: attention averages evidence over a window set by its context length, whereas the SSM discounts past evidence at rates set by its recurrent transition. At first glance, one might expect full-context attention ($L = T$) to be optimal and the SSM's exponentially decaying kernel to forget useful evidence, since θ is fixed and no observation becomes stale. But linear attention's optimal kernel cannot adapt to elapsed time: at $L = T$, it weights every evidence atom by $\Theta(1/T)$, severely underweighting the early evidence and causing large cold-start bias. Balancing this against the variance from a shorter window then yields $L^\star(T) = \Theta(\sqrt{T})$, not T. This also provides a theoretical analogue of empirical findings that longer context windows need not improve LLMs' effective context use (Liu et al., 2024; Hsieh et al., 2024). Even at the optimal context length, attention pays a shape factor $2/\sqrt{3}$: the SSM's exponential kernel achieves a better bias–variance trade-off than attention's uniform kernel. Beyond this shape factor, the architectural comparison also depends on the task's statistical structure, namely how the signal variance is spread across eigendirections (the λ_i). The SSM tunes a separate decay rate to each eigendirection, whereas attention's single context length must compromise across all of them. Thus, at fixed noise and total signal variance, concentrating the signal in fewer eigendirections lowers ζ but leaves $\bar{\zeta}$ unchanged, increasing the SSM's advantage. App. E.1.6 further examines these statistical effects and also the loss curvature at the optimal weights; App. D.3.1 experimentally validates Prop. 1.

Stationary optima as inductive biases. Practical attention and SSMs are not confined to stationary kernels. Indeed, normalization and additional depth let both families escape the stationary regret floor (Appendix E.2). However, the stationary comparison remains consequential: it isolates the temporal weighting supplied by each family's core mechanism and predicts which kernels richer architectures in each family learn most readily under finite optimization—their inductive biases.

We test this hypothesis using single-layer instances of the practical architectures described in §2.1. We use two scalar Gaussian filtering tasks: one with a uniform and another with an exponential Bayes kernel. Fig. 2 compares each regime's Bayes kernel with the learned aggregation kernels estimated from model predictions. App. D.1.1 gives full experimental details.

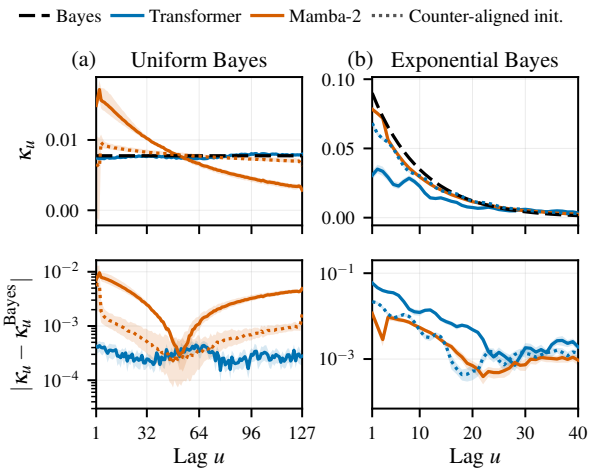


Figure 2: Kernel alignment in practical models under finite optimization. Shading shows 95% bootstrap CIs over 20 seeds.

Regardless of the regime, the Transformer learns flatter kernels, whereas Mamba-2 learns exponentially decaying ones, mirroring the in-class optimal kernels of Prop. 1. Correspondingly, the Transformer performs better in the uniform regime and Mamba-2 in the exponential regime. The mismatch is asymmetric under random initialization: Mamba-2 incurs $1.9\times$ the Transformer's regret in the uniform regime, whereas the Transformer incurs $4.3\times$ Mamba-2's in the exponential regime, echoing Proposition 1, where the exponential kernel attains the stationary optimum. To distinguish representational limitations from inductive bias, we also design a counter-aligned initialization (dotted lines) that biases each model toward the other Bayes kernel—uniform for Mamba-2 and exponential for the Transformer. Without changing either architecture, this initialization largely closes both performance gaps, showing that both models can closely approximate either kernel but more readily learn the one favored by their stationary core.

4 Case 2: Routing

With a nontrivial routing map $\rho(s) < s$, each label y_s is separated from the regressor $x_{\rho(s)}$ that generated it. This creates two architectural demands. First, the architecture must assemble each evidence atom $x_{\rho(s)}y_s$ from factors appearing in different tokens. Second, it must address the past regressor $x_{\rho(t+1)}$ to which the maintained belief is applied. Once an architecture satisfies these demands, the remaining problem is the Case 1 kernel comparison. Accordingly, two elements of belief geometry become central in Case 2: *belief alignment*, which measures agreement between the implicit and Bayes beliefs and therefore captures failures of evidence assembly, and *addressing fidelity*, which measures whether the prediction applies its implicit belief to the correct regressor. Together, they yield the following lower bound.

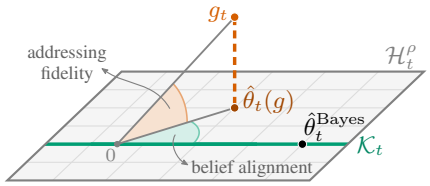


Figure 3: Floor factors in belief geometry.

Theorem 2 (Two-factor floor)**.** *For every g, with $\text{corr}^2_{\Sigma_x}$ denoting the squared cosine under $\langle\cdot,\cdot\rangle_{\Sigma_x}$,*

$$\text{Reg}_T(g) \;\geq\; \sum_{\substack{t<T:\\ t+1 \text{ matched}}} \|\hat{\theta}_t^{\text{Bayes}}\|^2_{\Sigma_x} \Bigg(1 - \underbrace{\text{corr}^2_{\Sigma_x}\left(\hat{\theta}_t(g), \hat{\theta}_t^{\text{Bayes}}\right)}_{\textit{belief alignment}} \cdot \underbrace{\frac{\|\hat{\theta}_t(g)\|^2_{\Sigma_x}}{\|\hat{\theta}_t(g)\|^2_{\Sigma_x} + \|g_t^{\perp}\|^2_t}}_{\textit{addressing fidelity}}\Bigg).$$

Both factors lie in $[0,1]$, so if either remains uniformly below 1 on a constant fraction of matched steps, then $\text{Reg}_T(g) = \Omega(T_\rho)$, where $T_\rho := |\{s \leq T : s \text{ matched}\}|$ is the *effective horizon* and we assume $T_\rho = \omega(\log T)$. The two routing rules below expose different sources of this floor: positional routing makes evidence assembly the bottleneck (Case 2.a, §4.1), whereas content routing makes addressing the bottleneck (Case 2.b, §4.2).

4.1 CASE 2.A: POSITIONAL ROUTING

Case 2.a introduces the positional routing map $\rho(s) = s - K$ for $s > K$, where $1 \le K < T$. The first K steps are unmatched, so $T_\rho = T - K$. Under this route, assembling each evidence atom $x_{s-K}y_s$ requires reaching K steps into the past. The same reach also supplies the regressor x_{t+1-K} to which the belief is applied. Thus, solving evidence assembly also solves addressing.

One might expect retaining the relevant history to suffice for positional assembly: attention keeps past tokens in its context, while an SSM compresses them into its recurrent state. But neither alone pairs each y_s with x_{s-K}: Case 2.a requires a compositional mechanism that first supplies positional reach and then uses it for assembly. Without this composition, increasing width or context length cannot make belief alignment in Thm. 2 approach 1, yielding the following floors for Case 1 architectures.

Corollary 1 (Assembly floors)**.** *Under the positional-routing setup above:*

(a) $\mathrm{Reg}_T(\mathcal{F}_{\mathrm{att}}(1, d, H, L)) = \Omega(T_\rho)$ *uniformly in* d, H, L.

(b) $\mathrm{Reg}_T(\mathcal{F}_{\mathrm{SSM}}(1, d, H, d_{\mathrm{conv}})) = \Omega(T_\rho)$ *uniformly in* $d, H \ge 1$ *and* $1 \le d_{\mathrm{conv}} \le K$ *with* K *fixed.*

Both families escape these floors by the same two-stage composition. First, they form the lag-shifted token $(x_{s+1-K}, y_s, x_{s-K})$, reducing Case 2.a to Case 1: two-layer attention uses positional encoding in its first layer, whereas the one-layer SSM uses a width-$(K+1)$ convolution. Then, they reuse their respective kernels from Prop. 1 over the effective horizon T_ρ, yielding the following regret rates.

Proposition 2 (Positional-routing pair)**.** *Under the positional-routing setup above, when* $H_{\mathrm{s}}d_{\mathrm{s}} < p$, $\mathrm{Reg}_T(\mathcal{F}_{\mathrm{SSM}}(1, d_{\mathrm{s}}, H_{\mathrm{s}}, d_{\mathrm{conv}})) = \Omega(T_\rho)$ *uniformly in* d_{conv}. *When* $H_{\mathrm{a}}d_{\mathrm{a}} \ge 2p$ *and* $H_{\mathrm{s}}d_{\mathrm{s}} \ge p$, *setting* $\widetilde{L}^\star := (K+1) \vee L^\star(T_\rho)$ *gives*

$$\mathrm{Reg}_T\big(\mathcal{F}_{\mathrm{att}}(2, d_{\mathrm{a}}, H_{\mathrm{a}}, \widetilde{L}^\star)\big) = O\big(K \vee \sqrt{T_\rho}\big), \qquad \mathrm{Reg}_T\big(\mathcal{F}_{\mathrm{SSM}}(1, d_{\mathrm{s}}, H_{\mathrm{s}}, K{+}1)\big) = \Theta\big(\sqrt{T_\rho}\big).$$

When $K = O(\sqrt{T_\rho})$, *the attention bound is also sharp:* $\mathrm{Reg}_T(\mathcal{F}_{\mathrm{att}}(2, d_{\mathrm{a}}, H_{\mathrm{a}}, \widetilde{L}^\star)) = \Theta(\sqrt{T_\rho})$.

Architectural lessons. Together, Cor. 1 and Prop. 2 add a lesson beyond Case 1's leading constants: the two families allocate resources for positional reach and belief maintenance differently. Attention uses the context length L for both jobs, so its construction takes $\widetilde{L}^\star$ to span both the routing lag K and Prop. 1's optimal window $L^\star(T_\rho)$. The SSM instead separates them: $d_{\mathrm{conv}} = K+1$ supplies only the positional reach, while its recurrent decay is tuned independently for belief maintenance as in Prop. 1. This has a memory consequence: the SSM stores only a p-dimensional recurrent state and K residual-stream vectors for convolution, whereas attention retains a growing history of $\widetilde{L}^\star = \Theta(K \vee \sqrt{T_\rho})$ such vectors. The growing context requirement is not specific to our two-layer construction:

Proposition 3 (Attention context requirement)**.** *Under the positional-routing setup above, for every fixed* D, $\mathrm{Reg}_T(\mathcal{F}^{\mathrm{sm}}_{\mathrm{att}}(D, d, H, L)) = O(\sqrt{T_\rho})$ *requires* $L = \Omega(\sqrt{T_\rho})$, *uniformly in* d, H.

Props. 2 and 3 show that, when $K = o(\sqrt{T_\rho})$, the SSM reaches the same regret rate with strictly less memory than fixed-depth attention with standard caching, regardless of width or softmax use (App. F.2.4). This identifies a task regime that favors SSMs: short-range positional assembly coupled with long-term belief maintenance, echoing Mamba's empirical advantage in Bondaschi et al. (2026).

Practical validation. App. D.3.2 validates these results directly within our analytically tractable classes, and App. D.2 confirms these architectural lessons for the practical models of §2.1 on a conjugate Beta–Bernoulli bandit. Here we test whether these lessons also extend to nonconjugate setups (App. B), using a delayed-feedback reward prediction task for a 30-arm contextual logistic bandit ($p = 30$, $K = 16$). Fig. 4 reports a Bayes-normalized performance score computed from cumulative excess log loss: 0 corresponds to a prior-only predictor that ignores the history, and 1 to Bayes. Throughout, we use 0.8 as the score threshold for high performance. Because our focus is representational capability, each cell reports the score of the best validation-selected run from a fixed sweep over training seeds and learning rates. App. D.1.2 details this choice and the full protocol.

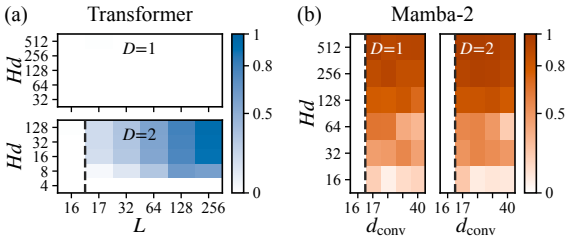
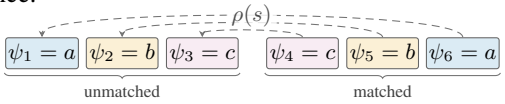


Figure 4: Practical validation of positional assembly mechanisms. Darker means closer to Bayes; dashed lines mark the $K + 1 = 17$ threshold.

Consistent with Cor. 1, the one-layer Transformer and one-layer Mamba-2 with $d_{\text{conv}} \leq K$ remain near baseline. Although depth extends the overall positional reach, both two-layer models improve only once their first layer can span the lag ($L, d_{\text{conv}} \geq K + 1 = 17$), matching the positional-assembly mechanism in Prop. 2. For the Transformer, $Hd \leq 8$ falls short, but once Hd reaches $16 < p = 30$, performance varies almost entirely with L, exceeding 0.8 at $L = 256$. This matches the architectural lesson that L, not Hd, governs belief maintenance while also supplying positional reach. For Mamba-2, $D = 2$ notably leaves the $d_{\text{conv}} = 17$ boundary unchanged, although two layers enable shorter convolutions for fixed-transition SSMs on the Gaussian task (App. D.3.2). At either depth, Mamba-2's performance improves primarily with Hd and first exceeds 0.8 at $Hd = 128$, well above $p = 30$. This fits the complementary lesson for SSMs: convolution supplies positional assembly, whereas recurrent-state capacity governs belief maintenance. Finally, the resource differences are reflected in total memory: the lowest-memory tested Transformer exceeding 0.8 uses $5.2\times$ as much as its Mamba-2 counterpart (App. D.1.3), echoing the theoretical memory comparison.

4.2 Case 2.b: content routing

Case 2.b introduces a fixed codebook of $R > 0$ unit vectors $\{\psi^{(1)}, \ldots, \psi^{(R)}\} \subset \mathbb{R}^{d_{\text{key}}}$. Writing ψ_s for the key observed at step s, we define the content routing map $\rho(s) := \max\{r < s : \psi_r = \psi_s\}$ for $s > R$. Following the layout in §2, each token additionally includes the current and next keys: $z_t = (\psi_{t+1}, x_{t+1}, y_t, \psi_t, x_t)$. Keys arrive in R-step cycles, each an independent uniform permutation of the codebook. The first cycle is unmatched, so $T_\rho = T - R$. Under this route, content rather than a fixed positional offset identifies the relevant past regressor, yielding a continuous-valued variant of associative recall, which is a widely studied sequence-model benchmark (Arora et al., 2023; Huang et al., 2025). In our case, however, it is embedded in Bayesian linear regression and must be solved alongside evidence assembly and belief maintenance.



The diagram to the right shows one realization of the first two cycles for $R = 3$, with codebook $\{a, b, c\}$. Its unequal spans illustrate the random lags $s - \rho(s) \in \{1, \ldots, 2R - 1\}$, so no fixed offset can locate all the previous occurrences. To avoid the positional-reach issue encountered in Case 2.a, we take $R = o(\sqrt{T_\rho})$, so $L^\star(T_\rho)$ spans every lag. At the same time, we keep content addressing nontrivial with a compressed codebook satisfying $d_{\text{key}} = \Theta(\log R)$ and $\max_{i \neq j} |\langle \psi^{(i)}, \psi^{(j)} \rangle| < 1/3$.

Here, unlike in Case 2.a, solving evidence assembly does not automatically solve addressing. Belief maintenance allows evidence assembly to tolerate coarse key matching: for $r \neq s$, the regressor $x_{\rho(r)}$ is independent of y_s and has zero mean, so spurious terms $x_{\rho(r)} y_s$ average out in the belief. But addressing cannot rely on this averaging: it requires the particular past regressor named by ψ_{t+1}. Thus, the architectures that suffice for Case 2.a can still make belief alignment in Theorem 2 approach 1, but they leave addressing fidelity bounded away from 1, forcing linear regret. Accordingly, we introduce *addressing capacity*, the central belief-geometry tool for quantifying this remaining failure.

Addressing capacity. Fix $t \geq R$. Let N_t be the number of keys not yet visited in the cycle containing $t + 1$, and index them by $q = 1, \ldots, N_t$. For each candidate q, let $X_q \in \mathbb{R}^p$ be the most recent regressor carrying that key, and let $g_{t,q}$ be the predictor's time-t output if the revealed key is q. The key idea behind addressing capacity is to decompose each candidate regressor into its components parallel and orthogonal to the Bayes-belief direction $\hat{\theta}_t^{\text{Bayes}}$. Let $\mathcal{C}_t$ collect the history with each candidate regressor replaced by its orthogonal component, leaving only their components along the Bayes-belief direction random (formalized in Appendix F). Crucially, conditioning on $\mathcal{C}_t$ makes the candidate Bayes predictions $(X_q^\top \hat{\theta}_t^{\text{Bayes}})_{q=1}^{N_t}$ independent Gaussians while leaving each $g_{t,q}$ free to depend arbitrarily on all of them through the candidate regressors' earlier occurrences. This allows us to decompose the analysis across Bayes-prediction directions and define

$$\text{Cap}_t(g) := \sum_{q=1}^{N_t} \text{corr}^2_{\mathcal{C}_t}\left(g_{t,q},\, X_q^\top \hat{\theta}_t^{\text{Bayes}}\right),$$

which is an analytically tractable measure of how many candidates are addressed. Because the arriving key is uniform, the average two-factor product in Theorem 2 is at most $\text{Cap}_t(g)/N_t$. Together with $N_t \geq R/2$ on half of every cycle, this bound implies that $\text{Cap}_t(g) = o(R)$ forces $\Omega(T_\rho)$ regret. Thus, deriving addressing floors reduces to upper-bounding addressing capacity, which we do next.

Theorem 3 (Addressing floors). *Under the content-routing setup above, fix $D \geq 1$ and $t \geq R$:*

(a) *Every $g \in \mathcal{F}_{\mathrm{att}}(D, d, H, L)$ satisfies $\mathrm{Cap}_t(g) = o(R)$, uniformly in d, H, L.*

(b) *Every $g \in \mathcal{F}^{\mathrm{aff}}_{\mathrm{SSM}}(D, d, H, d_{\mathrm{conv}})$ satisfies $\mathrm{Cap}_t(g) = o(R)$, uniformly in d, H, d_{conv}.*

(c) *Every $g \in \mathcal{F}^{\mathrm{sel}}_{\mathrm{SSM}}(D, d, H, d_{\mathrm{conv}})$ satisfies $\mathrm{Cap}_t(g) = O(DHd \log R)$, uniformly in d, H, d_{conv}.*

Thus, all predictors in (a)–(b) incur $\Omega(T_\rho)$ regret, as do those in (c) with $DHd = o(R/\log R)$.

Unlike prior bit-based lower bounds for associative recall and copying (Jelassi et al., 2024; Wen et al., 2024; Arora et al., 2025), Theorem 3 requires no finite-precision assumption and treats continuous, noisy prediction. The result also clarifies why softmax matters: linear attention is often motivated as a more efficient replacement for softmax attention (Katharopoulos et al., 2020; Schlag et al., 2021), yet here that replacement makes sublinear regret impossible at any fixed depth.

The following constructions escape the floors of Thm. 3 by replacing Case 2.a's fixed positional offset with *sharp nonlinear selection*. Attention selects from retained tokens with separate softmax heads for evidence assembly and addressing. The SSM instead encodes each candidate in two recurrent-state coordinates per head using a diagonal sigmoid gate for addressing and assembles evidence with a width-2 convolution. Both then reuse their Case 1 kernels from Prop. 1 for belief maintenance.

Proposition 4 (Content-routing pair). *Under the content-routing setup above,*

$$\mathrm{Reg}_T\big(\mathcal{F}^{\mathrm{sm}}_{\mathrm{att}}(2, d_{\mathrm{key}} \vee p, 2, L^\star(T_\rho))\big) = O(\sqrt{T_\rho}), \qquad \mathrm{Reg}_T\big(\mathcal{F}^{\mathrm{sel}}_{\mathrm{SSM}}(2, 2R, p, 2)\big) = O(\sqrt{T_\rho}).$$

Architectural lessons. Thm. 3 and Prop. 4 reveal a key architectural difference in how sharp selectivity is employed. Softmax attention selects among retained tokens, so it can address R candidate tokens with only $O(\log R)$-dimensional representations; the SSM instead preserves candidates separately in its recurrent state, so the first-layer-selective SSM class $\mathcal{F}^{\mathrm{sel}}_{\mathrm{SSM}}$ requires $\widetilde{\Omega}(R)$ recurrent-state dimensions, which our $O(R)$ construction matches up to logarithmic factors. Together, Cases 2.a–b yield a clean division of labor: convolution gives SSMs cheap positional reach, while softmax gives attention cheap content addressing.

Practical validation. We test two-layer practical models on Case 2.a's delayed-bandit task, replacing its fixed delay with content routing over a compressed codebook of size $R = 32$. Consistent with Theorem 3, Figure 5 shows that the no-softmax Transformer—despite retaining its SwiGLU MLPs—and ablated Mamba-2 with fixed transitions and no additional write nonlinearities—despite retaining nonlinear reads—remain far below Bayes. By contrast, the softmax Transformer and full Mamba-2 achieve much higher scores. For the Transformer, performance rises sharply when Hd reaches $12 < R = 32$ and then changes little with further width. Improvement with L is instead smooth, with no sudden jump from $L = 2R - 1 = 63$ to $L = 2R = 64$, reflecting Case 2.a's belief-maintenance trend without its positional-reach threshold. Full Mamba-2 instead depends mainly on Hd, with no systematic benefit from increasing d_{conv}.

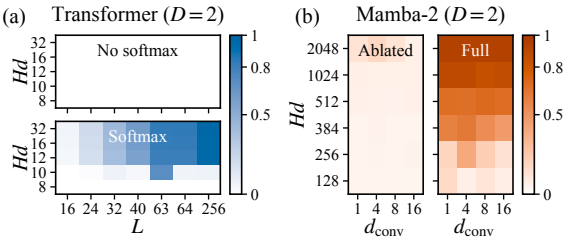


Figure 5: Practical validation of content addressing mechanisms. Darker means closer to Bayes.

Fig. 6 tracks the smallest tested Hd reaching 0.8 as R increases from 8 to 128. It stays within 10–24 for the Transformer but grows rapidly for Mamba-2 (hollow marker means 0.8 not reached). This qualitatively matches the exponential width separation jointly established by the lower bound in Thm. 3(c) and the constructions in Prop. 4. App. D.1.4 gives full experimental details.

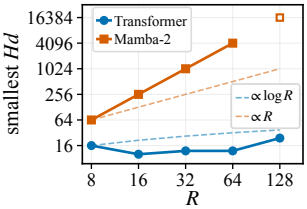


Figure 6: Width scaling (log–log).

5 Conclusion

In this paper, we developed belief geometry to compare the representational capabilities of attention and SSMs in ICLR. Because its elements abstract common computational demands, we expect them to aid analyses across broader settings. Another promising direction is to extend this representational framework to optimization, thereby helping explain why training favors some solutions over others.

AI USE STATEMENT

We used generative AI tools for refining theoretical claims and assisting with their proofs, implementing experiments, surveying related work, and revising the manuscript. We checked the mathematical arguments, experimental implementations, results, and cited sources, and take full responsibility for the final content of the paper. We did not use generative AI tools to generate research ideas or to propose hypotheses.

ETHICS STATEMENT

This work combines theoretical analysis of attention and SSMs with experiments on synthetic data. It involves no human participants or personal data. We identify no ethical concerns specific to the work presented.

REPRODUCIBILITY STATEMENT

Section 2.1 defines our architecture classes, with practical connections and extensions in Appendix C. Appendix D documents the experimental setups and training and evaluation protocols, while Appendices E and F provide full statements and proofs of the theoretical results. The accompanying anonymous code includes implementations, configurations, plotting data, and commands to rerun the experiments and recreate the figures.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. arXiv:2306.00297, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. arXiv:2211.15661, 2023.

Enes Arda, Semih Cayci, and Atilla Eryilmaz. Finite-Time Analysis of Gradient Descent for Shallow Transformers. In *AISTATS*, 2026.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and Improving Recall in Efficient Language Models. arXiv:2312.04927, 2023.

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:2402.18668, 2025.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection. arXiv:2306.04637, 2023.

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. arXiv:2405.04517, 2024.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a Transformer: A Memory Viewpoint. arXiv:2306.00802, 2023.

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*, pp. 6956–6967. PMLR, April 2023.

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. arXiv:2502.10178, 2026.

Semih Cayci and Atilla Eryilmaz. Convergence of Gradient Descent for Recurrent Neural Networks: A Nonasymptotic Analysis. arXiv:2402.12241, 2024.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training Dynamics of Multi-Head Softmax Attention for In-Context Learning: Emergence, Convergence, and Optimality. arXiv:2402.19442, 2024a.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling Induction Heads: Provable Training Dynamics and Feature Learning in Transformers. arXiv:2409.10559, 2024b.

Nicola Muca Cirone, Antonio Orvieto, and Benjamin Walker. Theoretical Foundations of Deep Selective State-Space Models. In *Neural Information Processing Systems*, 2024.

Edo Cohen-Karlik, Itamar Zimerman, Liane Galanti, Ido Atad, Amir Globerson, and Lior Wolf. On the Expressivity of Selective State-Space Layers: A Multivariate Polynomial Approach. arXiv:2502.02209, 2025.

Frank Cole, Yuxuan Zhao, Yulong Lu, and Tianhao Zhang. In-Context Learning of Linear Dynamical Systems with Transformers: Approximation Bounds and Depth-Separation. arXiv:2502.08136, 2025.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. arXiv:2405.21060, 2024.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models. arXiv:2402.19427, 2024.

Benjamin L. Edelman, Ezra Edelman, Surbhi Goel, Eran Malach, and Nikolaos Tsilivis. The Evolution of Statistical Induction Heads: In-Context Learning Markov Chains. arXiv:2402.11004, 2024.

Chanakya Ekbote, Marco Bondaschi, Nived Rajaraman, Jason D. Lee, Michael Gastpar, Ashok Vardhan Makkuva, and Paul Pu Liang. What One Cannot, Two Can: Two-Layer Transformers Provably Represent Induction Heads on Any-Order Markov Chains. arXiv:2508.07208, 2025.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved Optimistic Algorithms for Logistic Bandits. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3052–3060, 2020.

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. arXiv:2212.14052, 2023.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What Can Transformers Learn In-Context? A Case Study of Simple Function Classes. arXiv:2208.01066, 2023.

Gautam Goel and Peter Bartlett. Can a transformer represent a Kalman filter? In *Proceedings of the 6th Annual Learning for Dynamics & Control Conference*, pp. 1502–1512. PMLR, June 2024.

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:2411.12537, 2025.

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently Modeling Long Sequences with Structured State Spaces. arXiv:2111.00396, 2022.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? arXiv:2404.06654, 2024.

Ningyuan Huang, Miguel Sarabia, Abhinav Moudgil, Pau Rodriguez, Luca Zappella, and Federico Danieli. Understanding Input Selectivity in Mamba: Impact on Approximation Power, Memorization, and Associative Recall Capacity. arXiv:2506.11891, 2025.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-Context Convergence of Transformers. arXiv:2310.05249, 2023.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. Repeat After Me: Transformers are Better than State Space Models at Copying. arXiv:2402.01032, 2024.

Jiarui Jiang, Wei Huang, Miao Zhang, Taiji Suzuki, and Liqiang Nie. Trained Mamba Emulates Online Gradient Descent in In-Context Linear Regression. arXiv:2509.23779, 2025.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. arXiv:2006.16236, 2020.

Zhong Li and Jiequn Han. Approximation and Optimization Theory for Linear Continuous-Time Recurrent Neural Networks. *Journal of Machine Learning Research*, 2022.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. Jamba: A Hybrid Transformer-Mamba Language Model. arXiv:2403.19887, 2024.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Arvind Mahankali, Tatsunori B. Hashimoto, and Tengyu Ma. One Step of Gradient Descent is Provably the Optimal In-Context Learner with One Layer of Linear Self-Attention. arXiv:2307.03576, 2023.

Pierre Marion, Raphaël Berthier, Gérard Biau, and Claire Boyer. Attention layers provably solve single-location regression. arXiv:2410.01537, 2025.

William Merrill, Jackson Petty, and Ashish Sabharwal. The Illusion of State in State-Space Models. arXiv:2404.08819, 2025.

Alireza Mousavi-Hosseini, Clayton Sanford, Denny Wu, and Murat A Erdogdu. When Do Transformers Outperform Feedforward and Recurrent Networks? A Statistical Perspective. In *Neural Information Processing Systems*, 2025.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How Transformers Learn Causal Structure with Gradient Descent. arXiv:2402.14735, 2024.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context Learning and Induction Heads. arXiv:2209.11895, 2022.

Jongho Park, Jaeseung Park, Zheyang Xiong, Nayoung Lee, Jaewoong Cho, Samet Oymak, Kangwook Lee, and Dimitris Papailiopoulos. Can Mamba Learn How to Learn? A Comparative Study on In-Context Learning Tasks. arXiv:2402.04248, 2024.

Bo Peng, Ruichong Zhang, Daniel Goldstein, Eric Alcaide, Xingjian Du, Haowen Hou, Jiaju Lin, Jiaxing Liu, Janna Lu, Guangyu Song, Kaifeng Tan, Saiteja Utpala, Nathan Wilce, Johan S Wind, Tianyi Wu, Daniel Wuttke, and Christian Zhou-Zheng. RWKV-7 "Goose" with Expressive Dynamic State Evolution. arXiv:2503.14456, 2025.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield Networks is All You Need. arXiv:2008.02217, 2021.

Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A Tutorial on Thompson Sampling. *Foundations and Trends in Machine Learning*, 11:1–96, 2018. doi: 10.1561/2200000070.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational Strengths and Limitations of Transformers. In *Neural Information Processing Systems*, 2023.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. One-layer transformers fail to solve the induction heads task. arXiv:2408.14332, 2024a.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, Parallel Computation, and Logarithmic Depth. In *International Conference on Machine Learning*, 2024b.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear Transformers Are Secretly Fast Weight Programmers. arXiv:2102.11174, 2021.

Jerome Sieber, Carmen Amo Alonso, Alexandre Didier, Melanie N Zeilinger, and Antonio Orvieto. Understanding the Differences in Foundation Models: Attention, State Space Models, and Recurrent Neural Networks. In *Neural Information Processing Systems*, 2024.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (Learn at Test Time): RNNs with Expressive Hidden States. arXiv:2407.04620, 2025.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive Network: A Successor to Transformer for Large Language Models. arXiv:2307.08621, 2023.

Yudou Tian, Neeraj Mohan Sushma, Harshvardhan Mestha, Nicolo Colombo, David Kappel, and Anand Subramoney. Learning in the Recurrent State: Gradient Descent with Linear Recurrent Networks. arXiv:2410.11687, 2026.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. arXiv:1706.03762, 2017.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. arXiv:2212.07677, 2023.

Tomoya Wakayama and Taiji Suzuki. In-Context Learning Is Provably Bayesian Inference: A Generalization Theory for Meta-Learning. arXiv:2510.10981, 2025.

Ke Alexander Wang, Jiaxin Shi, and Emily B. Fox. Test-time regression: A unifying framework for designing sequence models with associative memory. arXiv:2501.12352, 2025.

Kaiyue Wen, Xingyu Dang, and Kaifeng Lyu. RNNs are not Transformers (Yet): The Key Bottleneck on In-context Retrieval. arXiv:2402.18510, 2024.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter L. Bartlett. How Many Pretraining Tasks Are Needed for In-Context Learning of Linear Regression? arXiv:2310.08391, 2024.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated Linear Attention Transformers with Hardware-Efficient Training. arXiv:2312.06635, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta Networks: Improving Mamba2 with Delta Rule. arXiv:2412.06464, 2025a.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing Linear Transformers with the Delta Rule over Sequence Length. arXiv:2406.06484, 2025b.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained Transformers Learn Linear Models In-Context. arXiv:2306.09927, 2023.

Yifan Zhang and Joe Kileel. Covering Number of Real Algebraic Varieties and Beyond: Improved Bounds and Applications. arXiv:2311.05116, 2025.

APPENDIX CONTENTS

Summary of theoretical results . . . 16
Guide to the proofs . . . 16
Notation guide . . . 17

A Extended related work . . . 18

B Beyond the linear-Gaussian testbed . . . 20

C Architecture classes: practical connections and extensions . . . 22
C.1 Connections to practical architectures . . . 22
C.2 Count-normalized extensions . . . 23

D Additional experiments and experimental details . . . 23
D.1 Details of the main-text experiments . . . 24
D.2 Validation on Beta–Bernoulli bandits . . . 29
D.3 Validation of the theoretical results . . . 30

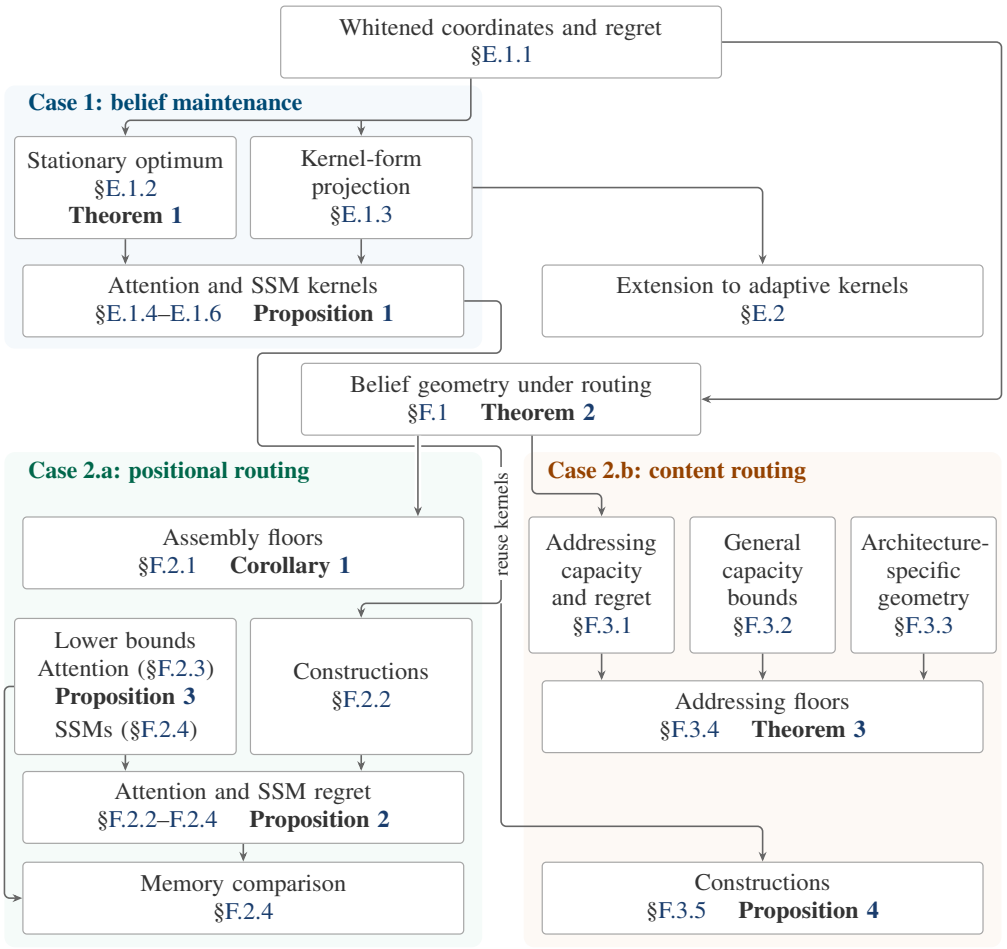
E Case 1: full statements and proofs . . . 33
E.1 Theorem 1 and Proposition 1: Stationary kernels . . . 33
E.2 Extension to time- and data-adaptive kernels . . . 45

F Case 2: full statements and proofs . . . 50
F.1 Proof of Theorem 2 . . . 50
F.2 Case 2.a: positional routing . . . 52
F.3 Case 2.b: content routing . . . 63

Table 1: Summary of theoretical results. Class notations are defined in §2.1.

	Bottleneck (Belief geometry tool)	Lower-bound classes	Construction classes
Case 1 (§3)	Belief maintenance (Aggregation kernel)	Proposition 1 $\mathcal{F}_{\mathrm{att}}(1,d,H,L)$, $Hd < p$ or fixed L $\mathcal{F}_{\mathrm{SSM}}(1,d,H,1)$, $Hd < p$	Proposition 1 $\mathcal{F}_{\mathrm{att}}(1,d,H,L^\star(T))$, $Hd \geq p$ $\mathcal{F}_{\mathrm{SSM}}(1,d,H,1)$, $Hd \geq p$
Case 2.a (§4.1)	Evidence assembly (Belief alignment)	Corollary 1 $\mathcal{F}_{\mathrm{att}}(1,d,H,L)$ $\mathcal{F}_{\mathrm{SSM}}(1,d,H,d_{\mathrm{conv}})$, $d_{\mathrm{conv}} \leq K$	Proposition 2 $\mathcal{F}_{\mathrm{att}}(2,d,H,\widetilde{L}^\star)$, $Hd \geq 2p$ $\mathcal{F}_{\mathrm{SSM}}(1,d,H,K+1)$, $Hd \geq p$
Case 2.b (§4.2)	Addressing (Addressing capacity)	Theorem 3 $\mathcal{F}_{\mathrm{att}}(D,d,H,L)$ $\mathcal{F}^{\mathrm{aff}}_{\mathrm{SSM}}(D,d,H,d_{\mathrm{conv}})$ $\mathcal{F}^{\mathrm{sel}}_{\mathrm{SSM}}(D,d,H,d_{\mathrm{conv}})$, $DHd = o(\frac{R}{\log R})$	Proposition 4 $\mathcal{F}^{\mathrm{sm}}_{\mathrm{att}}(2,d_{\mathrm{key}} \vee p,2,L^\star(T_\rho))$ $\mathcal{F}^{\mathrm{sel}}_{\mathrm{SSM}}(2,2R,p,2)$

Guide to the proofs. Boxes point to proof sections in the appendix, arrows show dependencies, and bold result numbers refer to the main text.



Notation guide. Frequently used notation, grouped by role. References point to the definitions.

Symbol	Meaning	Defined in
Data and routing		
$\theta,\ p$	Unknown regression parameter $\theta \in \mathbb{R}^p$ and its dimension.	§2
$x_s,\ y_s,\ z_s$	Regressor, response, and token $z_s = (x_{s+1}, y_s, x_s)$; Case 2.b also includes keys.	§2, §4.2
$\Sigma_x,\ \Sigma_\theta,\ \sigma^2$	Regressor covariance, prior covariance, and noise variance.	§2
$T,\ T_\rho$	Total time horizon; effective horizon (number of matched steps).	§2, §4
$\rho(s),\ K$	Regressor index paired with y_s; positional delay K gives $\rho(s) = s - K$ in Case 2.a.	§2, §4.1
$\mathcal{M}_t,\ N_t^\rho$	Matched observation indices through time t and their count.	§F.1
$\psi_s,\ R,\ d_{\text{key}}$	Content key, number of distinct keys, and key-vector dimension (Case 2.b).	§4.2
Predictions and belief geometry		
$g_t,\ g_t^{\text{Bayes}}$	Prediction of y_{t+1} from $z_{1:t}$, and its Bayes-optimal counterpart.	§2
$\hat{\theta}_t(g),\ \hat{\theta}_t^{\text{Bayes}}$	Implicit belief extracted from g_t by projection, and posterior mean $\mathbb{E}[\theta \mid z_{1:t}]$.	§2
$\text{Reg}_T(g),\ \text{Reg}_T(\mathcal{F})$	Cumulative excess squared loss over Bayes; the best achievable regret in class $\mathcal{F}$.	§2
$\mathcal{H}_t^\rho,\ g_t^\perp$	Belief-form prediction subspace; component of g_t orthogonal to it.	§2
$\mathcal{K}_t,\ \Pi_t$	Kernel-form belief subspace and orthogonal projection onto it.	§2
$\|\cdot\|_t,\ \|\cdot\|_{\Sigma_x},\ \text{corr}^2_{\Sigma_x}$	L^2 norm for predictions and whitened beliefs; Σ_x-weighted belief norm; belief alignment.	§2, §4, §F.1
$K_{t,s},\ \kappa_u$	Matrix-valued aggregation-kernel weight on $x_{\rho(s)}y_s$; stationary kernels have $K_{t,s} = \kappa_{t-s}$.	§2, §3
$N_t,\ X_q,\ g_{t,q}$	Number of candidate next keys, their latest regressors, and predictions under each candidate key (Case 2.b).	§4.2
$\mathcal{C}_t,\ \text{Cap}_t(g)$	Conditioning information making candidate Bayes predictions independent; capacity sums their squared conditional correlations with $g_{t,q}$.	§4.2
Architectural resources		
$D,\ H,\ d,\ d_e$	Depth, heads per layer, per-head width (SSM state width), and residual-stream width.	§2.1
$L,\ d_{\text{conv}}$	Attention window length and convolution tap count; a convolution reaches $d_{\text{conv}} - 1$ steps back.	§2.1
$L^\star(T),\ \widetilde{L}^\star$	Optimal Case 1 attention window; $\widetilde{L}^\star = (K+1) \vee L^\star(T_\rho)$ also covers positional reach.	§3, §4.1
Coordinates and statistics in the proofs		
$\widetilde{x}_t,\ \widetilde{\theta}$	Whitened variables: $\widetilde{x}_t = \Sigma_x^{-1/2}x_t$ and $\widetilde{\theta} = \Sigma_x^{1/2}\theta$.	Eq. (3)
$\widetilde{\Sigma}_\theta,\ U,\ \lambda_i$	Whitened prior covariance $\Sigma_x^{1/2}\Sigma_\theta\Sigma_x^{1/2}$, its eigenbasis, and its eigenvalues.	Eq. (3)
$\widetilde{\theta}_t(g),\ \widetilde{\theta}_t^{\text{Bayes}}$	Whitened beliefs: $\Sigma_x^{1/2}\hat{\theta}_t(g)$ and $\Sigma_x^{1/2}\hat{\theta}_t^{\text{Bayes}}$, respectively.	§E.1.1
$K_u,\ M_t$	Whitened stationary kernel $K_u = \Sigma_x^{1/2}\kappa_u\Sigma_x^{1/2}$ and kernel mass $M_t = \sum_{u<t} K_u$ (Case 1).	Eq. (9)
$\Gamma,\ v_i$	Evidence-innovation covariance and its eigenvalues $v_i = \lambda_i + \sum_j \lambda_j + \sigma^2$.	Eq. (4)
$\zeta,\ \bar{\zeta}$	Bias–variance scales: $\zeta = \sum_i \sqrt{\lambda_i v_i}$ and $\bar{\zeta} = \sqrt{(\sum_i \lambda_i)(\sum_i v_i)}$.	§3
$A_t,\ B_t$	Accumulated evidence $\sum_{s\le t} x_s y_s$ and accumulated information $\sum_{s\le t} x_s x_s^\top$ (Case 1).	§E.1.1
$j,\ n,\ Y_j,\ r_j$	Case 2.a: $j = s - K$, $n = t - K$; matched response $Y_j = y_{K+j}$ and evidence atom $r_j = \widetilde{x}_j Y_j$.	§F.2

A Extended related work

This appendix expands the comparisons of §1 result by result, following the three capabilities of our belief geometry, and then places our analysis relative to other expressivity axes, unifying frameworks, and optimization results.

Belief maintenance and in-context regression (Case 1). Our testbed descends from the linear-regression formulation of in-context learning of Garg et al. (2023), and the theory that followed asks which estimator a model can implement: transformers can be constructed to run gradient descent and ridge regression (Akyürek et al., 2023), one layer of linear self-attention performs one gradient step (von Oswald et al., 2023), and deeper constructions implement ridge, Lasso, and generalized linear estimators with in-context algorithm selection (Bai et al., 2023). Statistical treatments bound the risk at a fixed context length (Wu et al., 2024; Wakayama & Suzuki, 2025). On the recurrent side, the same gradient step has a one-layer recurrent implementation (Tian et al., 2026). Closest to belief maintenance itself, transformers can implement the Kalman filter and recurrent networks can approximate Bayesian filters, both with time-uniform error bounds (Goel & Bartlett, 2024; Bishop & Bonilla, 2023), while linear recurrences efficiently approximate a temporal relationship only when its dependence on the distant past decays sufficiently (Li & Han, 2022).

Prior work thus establishes implementability or bounds the loss at a fixed prediction step. Our belief-maintenance analysis instead asks what maintaining the Bayes belief costs over the stream, comparing both families under cumulative Bayes regret. Our stationary-floor theorem (Theorem 1) lower-bounds every stationary kernel by $\zeta\sqrt{T} + o(\sqrt{T})$, and Proposition 1 shows that once $Hd \geq p$, the SSM's exponential kernel attains this floor while attention's uniform kernel pays the factor $\frac{2}{\sqrt{3}}\,\bar{\zeta}/\zeta$. Among the closest architecture lower bounds, that of Cole et al. (2025) for a single-layer linear class is a non-vanishing terminal loss under a parameter-norm constraint; our analysis concerns cumulative regret and identifies the exact leading $\sqrt{T}$ constant.

Evidence assembly and positional routing (Case 2.a). Positional evidence assembly requires composing operations across tokens: locating the relevant past input and combining it with the current response. Related work examines how attention depth and convolution in SSMs enable such interactions. For attention, one layer cannot solve the induction-heads task unless its size grows with the sequence, while two layers suffice (Sanford et al., 2024a); two layers with one head each represent any conditional k-gram (Ekbote et al., 2025); a single nonlinear attention layer is asymptotically Bayes optimal for regression on one latent location (Marion et al., 2025); and logarithmic depth separates from a single layer on dynamical systems (Cole et al., 2025). On the recurrent side, positional reach comes from short convolutions: H3 introduced its shift operator because state-space models struggled to recall earlier tokens (Fu et al., 2023), and fixed convolution filters support only a subset of token-interaction distances (Arora et al., 2023). For Markov-chain prediction, Bondaschi et al. (2026) show that a simplified one-layer Mamba with width-2 convolution and an ℓ_1-normalized readout exactly represents the Bayes-optimal predictor for their first-order Markov model.

The cited attention results establish depth requirements and prediction guarantees for specific tasks, while the recurrent constructions identify how convolution enables useful predictors. Our analysis connects these mechanisms through a common cumulative-regret criterion. Corollary 1 quantifies the cost of insufficient evidence assembly as linear regret in both architecture classes. Beyond establishing that assembly is possible, Propositions 2 and 3 compare the memory needed for $\Theta(\sqrt{T_\rho})$ regret: the SSM separates local positional reach from belief maintenance, whereas fixed-depth attention in our classes requires a growing context to match this rate. Under standard per-position caching, this yields an SSM memory advantage when $K = o(\sqrt{T_\rho})$.

Content addressing (Case 2.b). Content addressing requires selecting the relevant information from a model's history. Induction heads illustrate this content-based selection: after observing $[A][B]$, they predict $[B]$ when $[A]$ reappears. Olsson et al. (2022) identified these circuits as a mechanism behind in-context learning. Mechanistic accounts describe associative retrieval through fast weights in linear attention (Schlag et al., 2021) and through modern Hopfield networks in softmax attention (Ramsauer et al., 2021). Arora et al. (2023) trace most of the observed language-modeling gap between attention and gated-convolution models to failures to recall earlier context and introduce multi-query associative recall to study this capability. Related experiments show that Mamba trails

transformers on retrieval while matching them on regression (Park et al., 2024). Hybrid models such as Griffin and Jamba combine recurrence with attention, improving retrieval and few-shot performance over their recurrent-only baselines (De et al., 2024; Lieber et al., 2024).

These empirical differences have theoretical counterparts. For sparse averaging, attention needs width logarithmic in the sequence while any recurrent model needs $\Omega(N)$ bits of state (Sanford et al., 2023). For k-hop induction, depth $\lfloor \log_2 k \rfloor + 2$ suffices, whereas recurrent models need depth k or polynomial width, with similar bounds for kernel-based and sparse attention (Sanford et al., 2024b). Two-layer transformers copy strings of exponential length while fixed-state models cannot (Jelassi et al., 2024); recurrent models with sublinear memory cannot retrieve even with chain of thought (Wen et al., 2024); and any recurrent model needs $\Omega(N)$ bits for exact associative recall (Arora et al., 2025). Mousavi-Hosseini et al. (2025) also establish a sample-complexity advantage for attention over recurrent networks on sparse-token regression.

Our addressing-capacity analysis differs from these results in three respects. First, at unbounded precision, the bit-based memory bounds above are vacuous as state-dimension bounds and do not distinguish transition mechanisms; Theorem 3 instead bounds real-valued state dimension with no finite-precision assumption and separates the mechanisms at fixed depth: linear attention incurs linear regret at any width and context length, fixed and affinely gated transitions incur it at every state width, sigmoid-gated transitions require $DHd = \Omega(R/\log R)$, and our sigmoid-gated construction succeeds with $\Theta(R)$ recurrent-state dimension (Proposition 4). Second, the copying and associative-recall tasks above are standalone and discrete, scored by exact success; our addressing problem is embedded in noisy Bayesian inference, alongside evidence assembly and belief maintenance, and scored by regret. Third, sparse averaging and sparse-token regression hand the model the relevant positions explicitly, whereas our key must be matched against R near-orthogonal codewords of width $\Theta(\log R)$, the axis on which softmax attention and sigmoid-selective SSMs separate.

Expressivity on other axes. Other expressivity results ask which sequential computations an architecture can perform, such as tracking a changing state or computing parity—whether a binary sequence contains an odd or even number of ones. S4, Mamba, and transformers lie in TC^0 and cannot solve general state-tracking tasks (Merrill et al., 2025). The transition spectrum also matters: linear recurrences whose transition matrices have only positive eigenvalues cannot compute parity, whereas allowing negative eigenvalues makes this possible (Grazzi et al., 2025). Gating provides another source of expressive power, enabling selective SSMs to represent functions beyond the capabilities of linear transformers (Cirone et al., 2024; Cohen-Karlik et al., 2025). These results characterize computational capabilities such as state tracking and parity. Our analysis instead quantifies the cumulative-regret and resource costs of assembling evidence, maintaining beliefs, and addressing past information.

Unifying frameworks. A complementary approach to comparing attention and SSMs is to identify the mathematical structure they share. Linear attention admits a recurrent form (Katharopoulos et al., 2020), and selective SSMs and attention are dual under structured state-space duality (Dao & Gu, 2024). The dynamical-systems framework of Sieber et al. (2024) rewrites attention, SSMs, and RNNs as one recurrence and identifies S6's input projections with attention's keys and queries. Test-time learning provides another common description: Sun et al. (2025) interpret sequence-modeling layers as test-time learners, recovering linear and softmax attention as special cases, while Wang et al. (2025) derive these layers as test-time regressors, with softmax attention as kernel smoothing.

These frameworks unify architectural representations. Our belief geometry instead unifies their analysis, translating architectural differences into task-dependent regret and resource separations. Appendix C details how the corresponding components of the practical architectures enter our formal classes.

Optimization. Our analysis is representational: it asks what each class can realize, not what gradient descent finds. The optimization literature is complementary, showing that related mechanisms are learnable in simplified settings. For linear attention, the one-step estimator is the global minimizer of the pretraining loss (Mahankali et al., 2023) and gradient flow reaches it (Zhang et al., 2023); softmax attention converges in phases and allocates heads to tasks (Huang et al., 2023; Chen et al., 2024a); induction heads emerge through distinct training phases (Bietti et al., 2023; Nichani et al., 2024; Edelman et al., 2024; Chen et al., 2024b); and Mamba trained by gradient descent implements

a variant of online gradient descent on linear regression (Jiang et al., 2025). Optimization guarantees also reveal a trade-off with memory: for diagonal recurrent networks, long-term dependencies can make gradient-descent guarantees grow exponentially with the sequence length (Cayci & Eryilmaz, 2024), whereas for shallow attention the optimization error is independent of the sequence length, at the price of memory that grows with it (Arda et al., 2026), echoing, from an optimization perspective, the memory trade-off we establish in Case 2.a.

B BEYOND THE LINEAR-GAUSSIAN TESTBED

Our three-operation viewpoint (evidence assembly, belief maintenance, addressing) in §2 extends beyond linear-Gaussian regression, and so does the belief geometry behind our regret analysis. Three examples below show how the operations persist as posterior updates and prediction rules change. We then identify when kernel form, the Pythagorean decomposition, and the two-factor floor carry over directly, and sketch how approximations can extend them further.

General formulation. Retain the online token layout and partial-injective routing of §2, including supplied keys, and evaluate predictors by the same squared-loss Bayes regret. Allow an arbitrary prior q_0 on θ and an arbitrary iid regressor distribution, with the parameter, regressors, and routing mutually independent. Responses are conditionally independent given these quantities, with density or mass function $p_\theta(y_s \mid x_{\rho(s)})$ and finite second moments.

Write q_t for the posterior distribution of θ given $z_{1:t}$ and $m_\theta(x) := \mathbb{E}[y_s \mid x_{\rho(s)} = x, \theta]$. Bayes' rule and conditional expectation give, at matched steps,

$$q_t(d\theta) \propto p_\theta(y_t \mid x_{\rho(t)}) q_{t-1}(d\theta), \qquad g_t^{\mathrm{Bayes}} = \int m_\theta(x_{\rho(t+1)})\, q_t(d\theta).$$

Unmatched observations leave the posterior unchanged, and unmatched targets have prediction zero.

Evidence assembly forms the likelihood contribution $p_\theta(y_t \mid x_{\rho(t)})$ from the matched pair $(x_{\rho(t)}, y_t)$. Belief maintenance updates q_{t-1} to q_t and preserves the posterior or a sufficient summary for subsequent updates and predictions. Addressing locates $x_{\rho(t+1)}$ before its response is observed, so that the maintained belief is applied to the correct regressor when forming g_t^{Bayes}.

Conjugate families. Let $\{q_\eta\}$ be a finite-dimensional conjugate exponential family with natural parameter η. In natural parameters, multiplying the prior by a likelihood factor adds the likelihood's contribution to the prior parameters. Denote this complete sufficient-statistic increment by $S(x, y)$. Starting from $q_0 = q_{\eta_0}$, conjugacy gives

$$q_t = q_{\eta_t}, \qquad \eta_t = \eta_{t-1} + S(x_{\rho(t)}, y_t) = \eta_0 + \sum_{s \le t} S(x_{\rho(s)}, y_s).$$

Here and below, sums include only matched steps, and the recursive update applies when t is matched. This connects directly to our kernel form: the natural parameters η_t aggregate the sufficient-statistic increments $S(x_{\rho(s)}, y_s)$. Evidence assembly forms $S(x_{\rho(t)}, y_t)$, belief maintenance accumulates η_t, and addressing supplies $x_{\rho(t+1)}$ for prediction under q_{η_t}. For Gaussian regression, $S(x, y) = (\sigma^{-2} xy, \sigma^{-2} xx^\top)$: the xy term is our evidence atom, while $xx^\top$ contributes to posterior precision and enters the Bayes kernel in Equation (1).

The following three examples make these operations explicit and the final paragraph discusses when and how our regret analysis extends.

Example 1 (conjugate): Beta–Bernoulli regression. Let the prior coordinates $\theta_j \sim \mathrm{Beta}(1, 1)$ be independent, and take iid regressors uniform on the standard basis vectors $e_1, \ldots, e_p$. Keep the same routing and let

$$y_s \mid \theta, x_{\rho(s)} \sim \mathrm{Bern}(x_{\rho(s)}^\top \theta).$$

With the accumulated statistics $N_t := \sum_{s \le t} x_{\rho(s)}$ and $A_t := \sum_{s \le t} x_{\rho(s)} y_s$, the centered Bayes belief has the kernel form

$$\hat{\theta}_t^{\mathrm{Bayes}} - \tfrac{1}{2}\mathbf{1} = \sum_{s \le t} K_{t,s}^{\mathrm{Bayes}} x_{\rho(s)} (y_s - \tfrac{1}{2}), \qquad K_{t,s}^{\mathrm{Bayes}} := \mathrm{diag}\left(\frac{1}{2 + N_{t,j}}\right)_{j=1}^{p},$$

where $\mathbf{1}$ is the all-ones vector. And the Bayes predictor is

$$g_t^{\mathrm{Bayes}} = x_{\rho(t+1)}^\top \hat{\theta}_t^{\mathrm{Bayes}}.$$

The three operations therefore have the same structure as in our linear-Gaussian setting. Evidence assembly forms the centered evidence atoms $x_{\rho(s)}(y_s - \frac{1}{2})$. Belief maintenance aggregates them through $K_{t,s}^{\mathrm{Bayes}}$ to obtain the centered Bayes belief $\hat{\theta}_t^{\mathrm{Bayes}} - \frac{1}{2}\mathbf{1}$. Addressing supplies $x_{\rho(t+1)}$ for prediction. Thus, this non-Gaussian example retains all three operations, the kernel form, and linear prediction form, although with different evidence atoms and a different aggregation kernel. App. D.2 tests our architectural lessons on this model under the Case 2 routings.

Example 2 (conjugate): lognormal regression. Take $p = 1$, a prior $\theta \sim \mathcal{N}(0, \nu)$ with $\nu > 0$, and iid regressors uniform on $[-1, 1]$. Keep the same routing and let

$$\log y_s = x_{\rho(s)}\theta + \epsilon_s, \qquad \epsilon_s \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

with $\sigma^2 > 0$ and the noises independent of the parameter, regressors, and routing. Completing the square gives $q_t = \mathcal{N}(\hat{\theta}_t^{\mathrm{Bayes}}, v_t)$, where $v_t^{-1} = \nu^{-1} + \sigma^{-2}\sum_{s\le t} x_{\rho(s)}^2$. The Bayes belief has kernel form

$$\hat{\theta}_t^{\mathrm{Bayes}} = \sum_{s\le t} K_{t,s}^{\mathrm{Bayes}} x_{\rho(s)} \log y_s, \qquad K_{t,s}^{\mathrm{Bayes}} := \sigma^{-2} v_t.$$

This is the same scalar Bayes kernel as in Equation (1), now applied to transformed evidence atoms. Evidence assembly forms different evidence atoms $x_{\rho(s)} \log y_s$. Belief maintenance aggregates them through $K_{t,s}^{\mathrm{Bayes}}$ to obtain $\hat{\theta}_t^{\mathrm{Bayes}}$. Averaging over the posterior and the next observation noise gives

$$g_t^{\mathrm{Bayes}} = \exp\left(x_{\rho(t+1)}\hat{\theta}_t^{\mathrm{Bayes}} + \frac{\sigma^2 + x_{\rho(t+1)}^2 v_t}{2} \right).$$

Addressing still supplies $x_{\rho(t+1)}$ for prediction. Thus, the three operations and the kernel form are retained but the Bayes prediction is a nonlinear function of the belief parameters.

Example 3 (nonconjugate): logistic regression. Keep the prior $\theta \sim \mathcal{N}(0, \nu)$, regressors, and routing of Example 2, but let matched responses $y_s \in \{0, 1\}$ be Bernoulli with $p_\theta(1 \mid x) = (1 + e^{-x\theta})^{-1}$. Multiplying the likelihood contributions gives

$$q_t(d\theta) \propto \exp\left\{ \theta \sum_{s\le t} x_{\rho(s)} y_s - \sum_{s\le t} \log(1 + e^{x_{\rho(s)}\theta}) \right\} q_0(d\theta).$$

Here, evidence assembly still combines each y_s with its generating regressor $x_{\rho(s)}$. Belief maintenance incorporates both the sum $\sum_{s\le t} x_{\rho(s)} y_s$ and the regressor-dependent logarithmic terms. Unlike Examples 1 and 2, this update does not directly yield a kernel-form Bayes belief. Addressing supplies $x_{\rho(t+1)}$, and averaging $p_\theta(1 \mid x_{\rho(t+1)})$ over q_t gives g_t^{Bayes}. The three operations remain, even though the Gaussian posterior update does not. The main-text Case 2 experiments use a p-dimensional version of this model (App. D.1).

Extending our regret analysis. The three operations apply to every prior and likelihood in the general formulation above. How much of our regret analysis extends with them depends on the structure of the Bayes prediction and of the posterior update.

When the Bayes prediction is linear in the Bayes belief through a fixed, finite-dimensional feature map φ of the addressed regressor, $g_t^{\mathrm{Bayes}} = \varphi(x_{\rho(t+1)})^\top \hat{\theta}_t^{\mathrm{Bayes}}$, as in Example 1 with $\varphi(x) = x$, the projection onto the belief subspace (Figure 1) extends directly, with $\varphi(x)$ in place of x and $\mathbb{E}[\varphi(x)\varphi(x)^\top]$ in place of Σ_x. So does the two-factor floor of Theorem 2, which rests on this projection alone.

The kernel-form refinement depends instead on the posterior update. When the posterior admits a finite-dimensional sufficient statistic that accumulates additively, as in conjugate families, belief

maintenance reduces to aggregating assembled increments, which mirrors the kernel form of Equation (1) and is exactly what happens in Examples 1 and 2. With both properties, as in Example 1, the full Pythagorean decomposition in Equation (2) extends directly, and with it our regret-analysis framework.

Beyond these conditions, kernel form (1) and the decomposition (2) still provide a starting point for further analysis. A nonlinear prediction without such features, as in Example 2, can be treated as a separate readout of a kernel-form belief. Linearizing the readout around the Bayes belief parameters makes the excess loss locally quadratic in the belief mismatch, and our projection decomposition then holds in the sensitivity-weighted geometry. A general posterior update, as in Example 3, can be approximated by one that retains kernel aggregation. Expanding the logarithmic terms to second order around the prior mean $\theta = 0$ gives a Gaussian approximation whose mean has the kernel form of Equation (1), with $\sigma^2 = 4$ and pseudo-responses $4(y_s - \frac{1}{2})$ in place of y_s. Combined with the readout linearization, the decomposition then applies to this approximate predictor, up to the errors of the posterior approximation and the readout linearization.

C Architecture classes: practical connections and extensions

Section 2.1 defines our architecture classes. Here we explain how their components connect to practical architectures and specify the count-normalized architecture extensions used in the adaptive results of Appendix E.2.

C.1 Connections to practical architectures

Table 2 collects the architectural references discussed below.

Table 2: Architectural references.

Architecture	Reference
Transformer	Vaswani et al. (2017)
LLaMA	Touvron et al. (2023)
S4	Gu et al. (2022)
H3	Fu et al. (2023)
RetNet	Sun et al. (2023)
Mamba	Gu & Dao (2024)
Mamba-2	Dao & Gu (2024)
mLSTM	Beck et al. (2024)
GLA	Yang et al. (2024)
DeltaNet	Yang et al. (2025b)
Gated DeltaNet	Yang et al. (2025a)
RWKV-7	Peng et al. (2025)

Attention. Our attention classes retain the Transformer's standard query–key scores, value aggregation, multi-head structure, output projections, and residual connections. The additive positional bias p_{t-s} is a tractable surrogate for the relative-lag dependence induced by RoPE in LLaMA-type Transformers.

SSMs: convolution, writes, and reads. Our SSM convolution corresponds to the short causal convolution that Mamba applies before its recurrence. Multiplicative writes and reads are shared by H3 and Mamba: writes form products of input features, while reads combine the stored state with current input features. Our formulation retains these multiplicative write/read operations, using affine feature projections and omitting additional nonlinear processing after the state read. Our vector-state formulation with coordinatewise multiplication also accommodates outer-product writes and matrix–vector reads. Indeed, index the recurrent state coordinates by matrix entries, $h_{t,(i,j)} = M_{t,ij}$. Repeating the affine projection rows gives

$$u_{t,(i,j)} = v_{t,i} k_{t,j}, \qquad o_{t,(i,j)} = M_{t,ij} q_{t,j}.$$

The output projection sums over j, recovering $(M_t q_t)_i$, with one recurrent coordinate per matrix entry.

SSMs: transitions. Fixed transitions support retention and decay, as in S4, H3, and RetNet. Scalar gates control retention across a head, as in Mamba-2's scalar-identity transition and mLSTM's scalar forget gate. Diagonal gates allow coordinatewise retention, as in Mamba and GLA. The delta rule makes a rank-one correction along the current key, following DeltaNet. Gated DeltaNet combines decay with such a correction. Our generalized delta form follows RWKV-7's diagonal-plus-rank-one update, allowing separate input and output directions. Our classes retain these transition mechanisms, using affine projections to form gate inputs and rank-one update vectors.

Our experiments use the practical architectures described in §2.1, retaining RMSNorm in both families, SwiGLU MLPs in the Transformer, and additional write/read nonlinearities in Mamba-2.

C.2 Count-normalized extensions

We use count normalization as a tractable surrogate for softmax normalization in attention and gated normalization in recurrent architectures such as mLSTM. It scales accumulated evidence by a deterministic token count, while learned parameters remain fixed throughout each stream. We use these extensions for the adaptive kernel constructions in Appendix E.2 and show that the assembly and addressing floors persist under count normalization in Appendix F. We specify a deterministic layerwise schedule $\mathbf{c} = (c_t^{(1)}, \dots, c_t^{(D)})$. For $D = 1$, we write c_t instead of $\mathbf{c}$.

Attention. Let $\tau_t := |J_t| = \min(t, L)$. Each layer uses $c_t^{(r)} \in \{1, \tau_t\}$ and replaces its linear weights by

$$a_{t,s}^{(r,b)} := \frac{s_{t,s}^{(r,b)}}{c_t^{(r)}}.$$

These weights define $\mathcal{F}_{\mathrm{att}}^{\mathrm{cal}}(D, d, H, L; \mathbf{c})$. The softmax-enabled class $\mathcal{F}_{\mathrm{att}}^{\mathrm{sm,cal}}(D, d, H, L; \mathbf{c})$ additionally permits any layer with $c_t^{(r)} \equiv 1$ to use softmax weights.

State-space models. Each layer uses $c_t^{(r)} \in \{1, t\}$, and only its ordinary read changes:

$$o_{t,\mathbf{c}}^{(r,b)} := \frac{(W_Q^{(r,b)} \widetilde{e}_t^{(r,b)}) \odot h_t^{(r,b)}}{c_t^{(r)}}.$$

Using these reads defines $\mathcal{F}_{\mathrm{SSM}}^{\mathrm{cal}}(D, d, H, d_{\mathrm{conv}}; \mathbf{c})$ for fixed transitions and $\mathcal{F}_{\mathrm{SSM}}^{\mathrm{sel,cal}}(D, d, H, d_{\mathrm{conv}}; \mathbf{c})$ for selective-first, fixed-later transitions. The same read-only normalization applies to $\mathcal{F}_{\mathrm{SSM}}^{\mathrm{aff}}$.

For every class, the all-one schedule $\mathbf{c} = \mathbf{1}$ recovers its base counterpart exactly. Count normalization adds no learned parameters and leaves recurrent-state dimension, residual width, and access to past tokens unchanged. It normalizes by a deterministic count, unlike the data-dependent normalization in softmax, LayerNorm, or RMSNorm.

D Additional experiments and experimental details

We first give full protocols and supplementary results for the practical experiments reported in the main text (§D.1). We then confirm the Case 2 architectural lessons for the same practical models on conjugate Beta–Bernoulli bandits (§D.2), complementing the nonconjugate tasks of the main text. Finally, we present additional experiments that test the exact theoretical predictions within our analytically tractable classes (§D.3).

Code availability. The code, experiment configurations, plotting data, and reproduction instructions are available in an anonymous repository:

https://anonymous.4open.science/r/belief-routing-experiments-2EF4/

D.1 Details of the main-text experiments

This section details the practical experiments reported in the main text. Section D.1.1 examines kernel alignment in Case 1 (§3). Section D.1.2 introduces the common setup and evaluation protocol for the Case 2 heatmaps. We then cover positional assembly in §D.1.3 (Case 2.a, §4.1) and selectivity and width scaling in §D.1.4 (Case 2.b, §4.2).

D.1.1 Case 1: kernel alignment

We detail the kernel-alignment experiments in §3 (Figure 2).

Tasks and tokens. We test whether practical models more readily learn the Bayes kernel favored by their stationary counterparts. Each episode contains $T = 128$ scalar observations, preceded by a single ordinary zero token. The model receives the scalar sequence $0, o_1, \dots, o_T$. The initial zero makes averaging the first $t+1$ tokens yield $\sum_{s=1}^{t} o_s/(t+1)$, the posterior mean in the uniform regime.

In the uniform regime, the posterior mean is $\hat{\theta}_t^{\mathrm{Bayes}} := \mathbb{E}[\theta \mid o_{1:t}]$:

$$\theta \sim \mathcal{N}(0,1), \qquad o_t = \theta + \epsilon_t, \quad \epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad \hat{\theta}_t^{\mathrm{Bayes}} = \frac{\sum_{s=1}^{t} o_s}{t+1}.$$

In the exponential regime,

$$\theta_1 \sim \mathcal{N}(0,1), \qquad \theta_{t+1} = \theta_t + \xi_t, \quad \xi_t \overset{\text{iid}}{\sim} \mathcal{N}(0,0.1), \qquad o_t = \theta_t + \epsilon_t, \quad \epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,9).$$

The posterior mean $\hat{\theta}_t^{\mathrm{Bayes}} := \mathbb{E}[\theta_t \mid o_{1:t}]$ satisfies the exact recursion $\hat{\theta}_0^{\mathrm{Bayes}} = 0$ and $\hat{\theta}_t^{\mathrm{Bayes}} = 0.9\hat{\theta}_{t-1}^{\mathrm{Bayes}} + 0.1 o_t$. The terminal Bayes aggregation kernels are therefore $\kappa_u^{\mathrm{Bayes}} = 1/(T+1)$ and $\kappa_u^{\mathrm{Bayes}} = 0.1(0.9)^u$, respectively, for $0 \le u < T$. The model output g_t estimates the posterior mean at all 128 observation times in each episode.

Models and optimization. We use the practical counterparts described in §2.1. Both models have a single pre-norm layer with residual width 32 (`d_model`) and a scalar linear readout. The Transformer uses four heads of width 8, full context $L = 129$, RoPE with base 10,000 and unscaled positions, and an MLP of width 64. Mamba-2 uses expanded width 64 (`d_inner`), state width 16, head dimension 16, one state group, convolution width 4, and chunk size 64.

Every run uses 450 AdamW updates with fresh batches of 64 episodes, peak learning rate 10^{-3} with 5% linear warmup and cosine decay to zero, $\beta_2 = 0.95$, weight decay 0.1, default remaining optimizer arguments, and global gradient-norm clipping at 5. The loss is mean squared error between g_t and $\hat{\theta}_t^{\mathrm{Bayes}}$, averaged over episodes and observation times; the reported model is the final iterate. We use 20 paired seeds. For each task and seed, all architecture and initialization conditions receive the same procedural training batches and the same independent evaluation set of 8192 episodes.

Counter-aligned initialization. For each architecture and seed, the counter-aligned model starts from the same random weights as the default model and then receives the change below; all weights remain trainable.

For the Transformer, counter-alignment plants half the log-slope of the exponential Bayes kernel, $\frac{1}{2}\log(0.9)$ per lag, in every head's attention scores. Since RoPE has no lag bias, it does so by changing the query and key weights of one rotary frequency only along the mean normalized observation embedding. For Mamba-2, counter-alignment averages the base parameters with an analytic endpoint that approximates the uniform posterior mean $\sum_{s=1}^{t} o_s/(t+1)$, using existing recurrent coordinates to accumulate the observation sum and token count, including the initial zero. The count coordinate is scaled so that Mamba-2's RMSNorm approximately divides the accumulated sum by the count. For both models, counter-alignment has strength $1/2$, uses no fitted prediction target, and changes only the initialization, not the architecture.

Fitted aggregation kernels. Following the kernel-fitting approach in §D.3.1, we describe each trained model's predictions through

$$g_T(o_{1:T}) \approx \sum_{u=0}^{T-1} \kappa_u o_{T-u}.$$

For each task and training seed, the final figure uses $N = 2^{17}$ fresh fitting episodes and 8192 independent held-out episodes, shared across its three model conditions. We estimate the kernel by ridge regression with penalty 10^{-2}. Held-out R^2 measures agreement with the trained model's predictions. Across the three conditions, mean R^2 over twenty seeds ranges from 0.9969–0.9993 in the uniform regime and 0.9641–0.9823 in the exponential regime.

Figure 2 plots the 20-seed mean fitted kernel and pointwise 95% percentile bands from 2000 bootstrap resamples of the training seeds. Its lower panels show the mean of the seedwise absolute deviations from the exact Bayes kernel. The mean all-lag relative ℓ_2 kernel errors are 6.13%, 48.82%, and 20.58% in the uniform regime, and 68.05%, 13.27%, and 21.13% in the order default Transformer, default Mamba-2, and counter-aligned model. One counter-aligned Mamba-2 run (seed 7012) failed to train, widening the dotted uniform band at short lags; the median counter-aligned uniform error is 11.41%.

Prediction error. We evaluate the trained models directly, independently of the kernel fit:

$$\mathcal{E} := \frac{\mathbb{E}[\sum_{t=1}^{T}(g_t - \hat{\theta}_t^{\text{Bayes}})^2]}{\mathbb{E}[\sum_{t=1}^{T}(\hat{\theta}_t^{\text{Bayes}})^2]}.$$

For the default Transformer and Mamba-2, respectively, the twenty-seed mean $\mathcal{E}$ is 3.21×10^{-3} and 6.09×10^{-3} in the uniform regime, and 7.96×10^{-2} and 1.86×10^{-2} in the exponential regime. Ratios computed before rounding give the main-text factors 1.9× and 4.3×. The Transformer has the lower error in 19 of 20 seeds in the uniform regime, and Mamba-2 in all 20 in the exponential regime. Counter-alignment reduces mean $\mathcal{E}$ by 60.4% for Mamba-2 in the uniform regime, including the failed run, and 70.5% for the Transformer in the exponential regime, relative to their default initializations.

Effect of training budget. Longer training can narrow the exponential-regime gap: in a two-seed run with 1800 updates, the Transformer acquires the exponential kernel, whereas its uniform-regime advantage persists. This does not change the inductive-bias picture of §3: each model learns its favored kernel readily, within a small budget, and the other more slowly.

D.1.2 Shared setup and evaluation for Case 2 heatmaps

To test whether our architectural lessons extend to nonconjugate setups (App. B), the heatmap experiments in Cases 2.a and 2.b use delayed-feedback contextual logistic bandits (Faury et al., 2020), specified in each subsection. Training and evaluation use only matched steps, when delayed feedback arrives. The prediction loss is binary cross-entropy to the current arm's sampled reward, averaged over episodes and matched prediction times; the Bayes prediction is used only for evaluation.

Let $g_t^{\text{Bayes}}, g_t \in (0, 1)$ denote the Bayes and model predictions at step t within an episode. We report the Bayes-normalized performance score, summing over matched steps in all evaluation episodes:

$$S := 1 - \frac{\sum_{\text{episodes}} \sum_{t:\text{ matched}} \mathrm{KL}\Big(\mathrm{Bern}(g_t^{\text{Bayes}}) \,\|\, \mathrm{Bern}(g_t)\Big)}{\sum_{\text{episodes}} \sum_{t:\text{ matched}} \mathrm{KL}\Big(\mathrm{Bern}(g_t^{\text{Bayes}}) \,\|\, \mathrm{Bern}(1/2)\Big)}.$$

This is cumulative excess log loss normalized between the prior-only predictor ($S = 0$) and Bayes ($S = 1$); scores can be negative.

Capability focus and training aids. To isolate what each architecture can attain, each heatmap cell reports the performance of the best validation-selected candidate from the fixed search specified in each subsection, rather than average performance over random initializations. To the same end,

training may use aids that make these solutions easier to find: an initialization favoring the routing lag in Case 2.a and an auxiliary objective built from past observable history in Case 2.b, detailed in each subsection. The aids act only during training; evaluation uses the unmodified architectures and native inputs, so they do not change what a model can represent. A success therefore shows what the architecture can represent, and a failure despite the aids is stronger evidence than a failure without them. Within each mechanism comparison, both variants of a family receive identical aids, so the aids cannot account for the contrast between them. All aids were fixed in development pilots before the reported grids were trained, and all runs in a comparison train for the same number of updates. Base learning-rate grids differ by family and width and were fixed in development runs. Wherever the top rate wins in a reported cell scoring above 0.1, we extend the grid until the selected rate lies strictly inside it.

D.1.3 CASE 2.A: POSITIONAL ROUTING

We detail the positional-routing experiments in §4.1 (Figure 4), using the shared setup and evaluation protocol of §D.1.2.

Task. We test the same positional-assembly demands in practical models on a delayed contextual logistic bandit with $p = 30$ arms, horizon $T = 600$, and delay $K = 16$. Each episode draws $\theta \sim \mathcal{N}(0, 4I_p)$, one coefficient per arm. At each step, the arm j is uniform, its context c has a uniform sign and magnitude drawn from $\mathrm{Unif}[0.5, 1.5]$, and the reward is Bernoulli with success probability $(1 + e^{-c\theta_j})^{-1}$, as in the logistic model of App. B. At time t, the model receives the current arm and context and the reward generated at $t - K$, but not the source arm or its context; recovering them is the positional-assembly problem. The input encodes the arm as a 30-dimensional one-hot vector, together with the context and the signed reward. The Bayes predictor is the posterior probability of a reward for the current arm and context, given all rewards received so far. This posterior has no closed form, so we compute it numerically.

Models. We report the Transformer at $L \in \{16, 17, 32, 64, 128, 256\}$, with $Hd \in \{32, 64, 128, 256, 512\}$ at depth one and $Hd \in \{4, 8, 16, 32, 128\}$ at depth two. Models below $Hd = 32$ have one head, and the others use head width 32; each SwiGLU MLP has width $4Hd$. Queries and keys use RoPE with base $10{,}000$ and unscaled positions. Before training, we set the query and key weights of the first head in the first layer to target the task's delay $K = 16$.

For Mamba-2, we test depths one and two with per-layer $Hd \in \{16, 32, 64, 128, 256, 512\}$ and $d_{\mathrm{conv}} \in \{16, 17, 24, 32, 40\}$. The residual width (`d_model`) is 16, expanded width (`d_inner`) is 32, head dimension is 8, and state width is chosen so that the total recurrent-state dimension per layer equals the displayed Hd; at $Hd = 16$, the expanded width is 16 with state width one. Convolutions start as causal identities, except that when $d_{\mathrm{conv}} \geq 17$, every first-layer channel starts with equal taps $1/\sqrt{2}$ at the current token and at lag 16. Both hints are training aids in the sense of §D.1.2: they only set initial values, and all parameters remain trainable.

Training and evaluation. All runs use AdamW for $12{,}000$ updates with effective batch size 16, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01, gradient clipping at 5, and constant learning rates.

Each Transformer heatmap cell searches four random seeds and five learning rates. At widths at most 128, the rates are $\{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}, 10^{-2}\}$; wider models use smaller rates in the same progression, from $3 \cdot 10^{-5}$ to $3 \cdot 10^{-3}$ at width 256 and from 10^{-5} to 10^{-3} at width 512. Each seed/rate candidate first trains for 6000 updates at $L = 256$. Every target context then independently restores that candidate's weights and Adam state, then continues through update $12{,}000$ at its target L, without restarting the optimizer or changing the rate.

Similarly, each Mamba-2 heatmap cell searches four random seeds and the five rates used for Transformer widths at most 128, adding $3 \cdot 10^{-2}$ at $(D, Hd, d_{\mathrm{conv}}) = (1, 256, 17)$, where the best of these lies at the edge of the grid.

In every cell, we select the candidate whose final iterate scores best on a fixed 1024-episode validation set, and report that model's score on 2048 independent test episodes, which never influence selection.

Results. The one-layer Transformer remains near baseline, with test scores below .01. Figure 4 shows that two-layer attention benefits mainly from longer context, whereas Mamba-2 benefits mainly from wider recurrent states once its convolution spans the delay. For example, the two-layer Transformer at $Hd = 16$ scores near zero at $L = 16$ and .200 at $L = 17$, increasing to .895 at $L = 256$; $Hd = 4$ remains near baseline throughout. For Mamba-2, $d_{\text{conv}} = 16$ remains near the prior-only baseline at every tested per-layer state dimension. At $d_{\text{conv}} = 17$, the scores for $Hd = \{16, 32, 64, 128, 256, 512\}$ are

Hd	16	32	64	128	256	512
$D = 1$	.225	.458	.654	.782	.881	.971
$D = 2$	.265	.511	.567	.824	.924	.975

Across all convolution widths, both depths first reach $S \geq 0.8$ at $Hd = 128$, although depth one does so only at $d_{\text{conv}} = 32$; at the $d_{\text{conv}} = 17$ boundary, it needs twice that width, $Hd = 256$. Depth two scores higher at every width except $Hd = 64$.

Memory comparison. We count learned parameters and persistent inference buffers for one sequence at a common precision. Among tested models with $S \geq 0.8$, the lowest-storage Transformer has $(D, L, Hd) = (2, 256, 16)$ and uses $9313 + 16{,}384 = 25{,}697$ scalar entries for parameters and a standard key–value cache. Its Mamba-2 counterpart has $(D, d_{\text{conv}}, Hd) = (1, 17, 256)$ and uses $3837 + 256 + 816 = 4909$ entries for parameters, recurrent state, and convolution buffers, respectively. Their ratio gives the main-text factor $5.2\times$.

Optimization effects. Per-seed results additionally show how reliably training reaches each cell's best run. For the two-layer Transformer at $L = 256$, only two of four training seeds reach $S \geq 0.8$ at $Hd = 32$, whereas all four score approximately .93 at $Hd = 128$, with learning rates selected separately within each seed by validation. Additional resources need not always help optimization, however. For Mamba-2 at $D = 2$ and $Hd = 64$, increasing d_{conv} from 24 to 40 lowers the test score from .583 to .242. A longer filter can reproduce a shorter one by zeroing its additional coefficients, so this decline suggests optimization sensitivity under the fixed training budget, rather than reduced representational capacity.

D.1.4 Case 2.b: content routing

We detail the content routing experiments in §4.2 (Figures 5 and 6), using the shared setup and evaluation protocol of §D.1.2.

Task. We test the role and resource cost of selectivity in practical models on a 10-arm contextual logistic bandit with horizon $T = 256$, using the coefficient prior, contexts, rewards, and Bayes predictor of Case 2.a (§D.1.3). Keys occur in independently permuted cycles, and the feedback arriving with a key is the reward generated at its previous occurrence. A shared unit-norm codebook uses dimensions $d_{\text{key}} = \{6, 8, 11, 14, 20\}$ for $R = \{8, 16, 32, 64, 128\}$. We construct the codebook with maximum absolute pairwise inner product below $1/3$, shared across models. Each token contains the current one-hot arm, codebook key, and context, the arriving signed reward, and the preceding arm, key, and context; the arriving reward's source arm and context are not given; recovering them is the content-addressing problem.

Models. For the mechanism heatmaps (Figure 5), we use $R = 32$ and $d_{\text{key}} = 11$. The two-layer, two-head Transformer uses a SwiGLU MLP with hidden width $4Hd$ and sweeps linear versus softmax attention over $Hd \in \{8, 10, 12, 16, 32\}$ and $L \in \{16, 24, 32, 40, 63, 64, 256\}$. Both variants use RoPE with base 100,000 and positions divided by 64, without a positional initialization hint.

The Mamba-2 heatmaps use two layers, residual width 64 (`d_model`), expanded width 128 (`d_inner`), head dimension 8, $Hd \in \{128, 256, 384, 512, 1024, 2048\}$, and $d_{\text{conv}} \in \{1, 4, 8, 16\}$. They compare stock Mamba-2 (Full) with two variants. Ablated removes sharp selection from both transitions and writes: in both layers, its transitions are input-independent but trainable, and the SiLU activations on the two write branches are replaced by their first-order linearizations at zero, so its writes are bilinear, as in our SSM classes. Fixed transitions removes only transition selectivity: its decay is input-independent but trainable, while its writes match Full's. All variants retain nonlinear

reads, output gating, normalization, and residual connections, initialize convolutions as identities, and use state-channel prefixes of a shared $Hd = 4096$ initialization.

For width scaling (Figure 6), we vary $R \in \{8, 16, 32, 64, 128\}$. The Transformer is a two-layer, two-head softmax model with $L = 256$ and $Hd \in \{2, 4, 6, 8, 10, 12, 16, 24, 32, 48, 64\}$, with the same RoPE settings at every R. Mamba-2 uses two selective layers, residual width 64, head dimension 8, $d_{\text{conv}} = 1$, and $Hd \in \{128, 256, 384, 512, 1024, 2048, 4096, 8192, 16384\}$. Its expanded width is 128 and its state width varies. At $R = 8, 16$, we additionally test $Hd \in \{8, 16, 32, 64\}$ with state width one and expanded width Hd.

Training and evaluation. All runs use 12,000 AdamW updates with batch size 32, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01, gradient clipping at 5, and constant learning rates. Training follows a fixed two-phase curriculum. In the alignment phase, only the input projection and first layer train, on an auxiliary target that pairs each arriving reward with its source arm, key, and context from the observed history. In the reward phase, the prediction loss takes over with a fresh optimizer. The Transformer aligns for 3000 updates and then trains all parameters on reward for 9000; Mamba-2 aligns for 9000 updates, then freezes the aligned layers and trains the rest on reward for 3000. Development pilots selected these schedules: the Transformer trained best with a short alignment phase followed by joint training, and Mamba-2 with a longer alignment phase whose layers then stay frozen.

Each Transformer cell searches three random seeds and learning rates $\{10^{-4}, 3 \cdot 10^{-4}, 5 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}, 10^{-2}, 3 \cdot 10^{-2}, 10^{-1}\}$. For the heatmaps, each seed/rate candidate first trains for 6000 updates at $L = 256$. Each target context independently restores that candidate's weights and Adam state, then continues for 6000 updates at its target L. Width-scaling runs (Figure 6) use $L = 256$ throughout.

Mamba-2 trains each configuration from scratch, searching eight random seeds and learning rates $\{3 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}\}$, extended with 10^{-2} and $3 \cdot 10^{-2}$ in some cells.

For both families, we select the candidate whose final iterate scores best on a fixed 256-episode validation set, and report its score on 2048 independent test episodes. The width-scaling frontier is the smallest tested width whose selected validation score reaches $S \geq 0.8$.

Results. Figure 5 shows that softmax attention and full Mamba-2 escape the low scores of their restricted variants: softmax attention reaches $S \geq 0.8$ in $9/35$ cells and full Mamba-2 in $8/24$, whereas linear attention and ablated Mamba-2 reach it in none. Fixed transitions reaches $S \geq 0.8$ in the same $8/24$ cells as Full (Figure 7), so only Ablated, which lacks sharp selection in both transitions and writes, stays near baseline, consistent with Theorem 3. Whereas the construction in Prop. 4 escapes the floor through selective transitions, Mamba-2's nonlinear writes provide a second route to the sharp selection that our floors identify as necessary. Transformer performance improves mainly with context once Hd reaches 12, whereas Mamba-2 depends mainly on Hd, with no systematic benefit from increasing d_{conv}. At $Hd = 12$, scores rise smoothly with context, from .045 at $L = 16$ to .808 at both $L = 63$ and $L = 64$, and .951 at $L = 256$. The width-scaling frontier in Figure 6 is

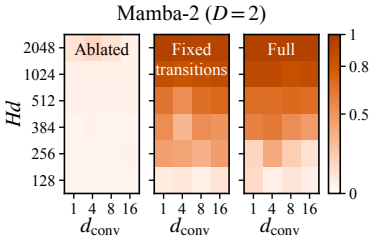


Figure 7: Content addressing with the three Mamba-2 variants ($R = 32$). Darker means closer to Bayes.

R	8	16	32	64	128
Transformer Hd	16	10	12	12	24
test score	.883	.933	.951	.946	.969
Mamba-2 Hd	64	256	1024	4096	–
test score	.874	.831	.889	.917	–

All nine frontier points exceed 0.8 on independent testing. At $R = 128$, no tested Mamba-2 width reaches $S \geq 0.8$; the best test scores are .283 at $Hd = 8192$ and .186 at $Hd = 16384$. Across the tested range, the successful Transformer width stays between 10 and 24, whereas the Mamba-2 frontier quadruples with each doubling of R, following $Hd = R^2$ from $R = 8$ to 64. This quadratic growth is consistent with Theorem 3(c), which only lower-bounds the required width by $\widetilde{\Omega}(R)$. It

suggests that practical Mamba-2 needs even more state than this floor, although part of this growth may reflect optimization under our fixed training budget. The dashed $\log R$ and linear-R curves are visual guides anchored at the first point, not fitted exponents.

Optimization effects. As in Case 2.a, architectural resources affect not only representational capability but also the outcomes of optimization. For the Transformer at $R = 128$, only one of three training seeds reaches $S \geq 0.8$ at the frontier width $Hd = 24$, whereas all three score .942–.974 at $Hd = 64$, with learning rates selected separately within each seed by validation. In Figure 5, the $Hd = 10$ row, just below the width where performance rises sharply, is irregular: .710 at $L = 63$ but .069 at $L = 64$ and .077 at $L = 256$. Since longer contexts only add visible tokens, this irregularity suggests fragile optimization at the threshold width rather than an effect of context length. For Mamba-2 at $R = 128$, the best test score falls from .283 at $Hd = 8192$ to .186 at $Hd = 16384$; since a wider state can reproduce a narrower one by leaving its extra channels unused, this decline again suggests optimization sensitivity under the fixed training budget, rather than a capacity limit.

D.2 Validation on Beta–Bernoulli bandits

We repeat the Case 2 experiments of §D.1 on conjugate Beta–Bernoulli bandits (Russo et al., 2018) with the same architectures and score, using earlier training recipes documented in the released code. Each episode draws independent arm means from $\mathrm{Unif}[0, 1]$, the $\mathrm{Beta}(1, 1)$ prior, and rewards are Bernoulli with the pulled arm's mean, as in the conjugate model of App. B. Case 2.a keeps $p = 30$ arms, $T = 600$, and $K = 16$, and Case 2.b keeps 10 arms, $T = 256$, and the same keys. The Bayes predictor is the exact Beta posterior predictive.

Case 2.a. Figure 8 reproduces the positional-assembly pattern of the main text. The one-layer Transformer stays near baseline, and both two-layer models improve only once their first layer spans the lag ($L, d_{\mathrm{conv}} \geq K + 1 = 17$). Two-layer attention then improves mainly with context and Mamba-2 mainly with recurrent-state width. The widths at which the models first exceed 0.8 differ from those on the nonconjugate task: $Hd = 8$ rather than 16 for the Transformer, and $Hd = 256$ rather than 128 for Mamba-2. The lowest-memory successful Transformer still stores 2.2× as many entries as its Mamba-2 counterpart.

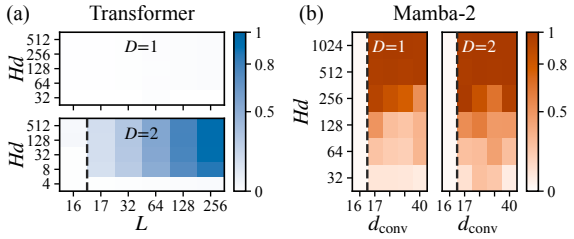


Figure 8: Positional assembly on a Beta–Bernoulli bandit ($p = 30$, $K = 16$). Darker means closer to Bayes; dashed lines mark the $K + 1 = 17$ threshold.

Case 2.b. Figures 9 and 10 reproduce the content-addressing pattern: softmax attention and full Mamba-2 escape the low scores of linear attention and ablated Mamba-2. At every R, both families need the same or smaller widths than on the nonconjugate task, and Mamba-2 now reaches 0.8 at $R = 128$, but the contrast is unchanged: the smallest successful width stays between 8 and 16 for the Transformer, whereas for Mamba-2 it again quadruples with each doubling of R, following $Hd = R^2/2$ from $R = 16$ to 128.

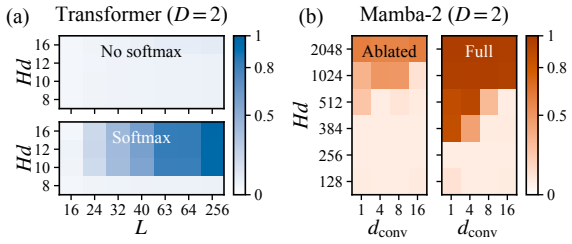


Figure 9: Content addressing on a Beta–Bernoulli bandit ($p = 10$, $R = 32$). Darker means closer to Bayes.

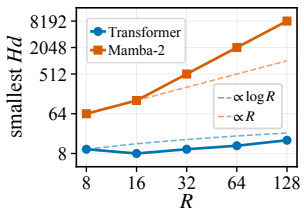


Figure 10: Smallest tested widths reaching $S \geq 0.8$ on validation.

D.3 Validation of the theoretical results

We use the architecture classes defined in §2.1 to test the predicted context scale and aggregation kernels in Case 1 (§D.3.1), depth and positional reach for evidence assembly in Case 2.a (§D.3.2), and selectivity for content addressing in Case 2.b (§D.3.3).

D.3.1 Case 1: belief maintenance

Context-length scaling. Proposition 1 predicts that the leading-order optimal window for one-layer linear attention is

$$L^\star(T) = \sqrt{3T \frac{\sum_i v_i}{\sum_i \lambda_i}}.$$

We test this prediction in the matrix-valued Gaussian Case 1 with

$$p = 2, \qquad \Sigma_x = I_2, \qquad \Sigma_\theta = \begin{bmatrix} 2.5 & -1.5 \\ -1.5 & 2.5 \end{bmatrix}, \qquad \sigma^2 = 1.$$

For this choice, $\sum_i \lambda_i = 5$ and $\sum_i v_i = 17$, giving $L^\star(T) = \sqrt{51T/5}$. We use $T \in \{512, 1024, 2048, 4096\}$ and sweep

$$\frac{L}{L^\star(T)} \in \left\{ \frac{1}{4}, \frac{1}{2}, \frac{1}{\sqrt{2}}, 1, \sqrt{2}, 2, 4 \right\},$$

rounding the resulting window to the nearest integer and plotting the realized ratio after rounding.

We train the one-layer linear-attention model $\mathcal{F}_{\text{att}}(1, 2, 1, L)$ with residual width $d_e = 6$, no positional bias, and a scalar readout. At time t, it receives $z_t = (x_{t+1}, y_t, x_t)$ and predicts y_{t+1}. All parameters are learned from a random initialization; no aggregation kernel is planted. We initialize all affine biases at zero and divide the attention output projection by $\sqrt{L}$ so that its initial scale is comparable across windows. Training minimizes squared error to the noisy response y_{t+1}, averaged over episodes and prediction times.

For each horizon–window pair (T, L), we use five paired random seeds and 5000 Adam updates per run, with initial learning rate 10^{-3}, cosine decay, and gradient clipping at 5. Each update uses a fixed batch of $2^{15} = 32{,}768$ freshly sampled tokens, with the number of episodes adjusted to the horizon. For a fixed horizon and seed, all window lengths receive the same training stream. We evaluate the final iterate on a common held-out set of 1024 episodes at each horizon, using cumulative squared-loss regret relative to the exact Bayes predictor.

Figure 11 shows the predicted U-shaped bias–variance trade-off. Normalizing by $\sqrt{T}$ approximately collapses the curves across horizons, and in every case the empirical valley is centered on the predicted scale. This location is not created by selecting the best run: the five-seed mean is also minimized by the $L/L^\star(T) = 1$ grid point at every horizon. Among the 20 individual seed–horizon profiles, 13 are minimized there and the remaining 7 at the adjacent $1/\sqrt{2}$ grid point; none is minimized elsewhere. Thus the experiment supports both the predicted $\sqrt{T}$ context scaling and the location of its leading-order optimum under finite optimization.

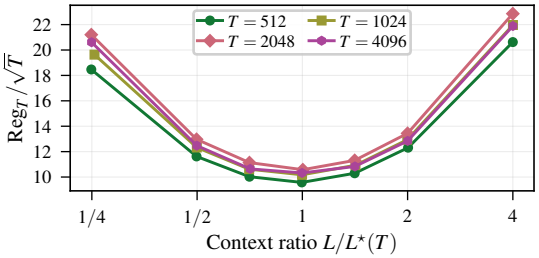


Figure 11: Context-length scaling in one-layer linear attention. Each point reports the minimum over five initializations.

Learned aggregation kernels. We next test whether training from random initialization recovers the kernels predicted by Proposition 1. At $T = 4096$, we reuse the five $\mathcal{F}_{\text{att}}(1, 2, 1, L^\star)$ runs at $L^\star(T) = 204$ and train five $\mathcal{F}_{\text{SSM}}(1, 2, 1, 1)$ runs. Both models have residual width 6 and use the 5000-update protocol above; no target kernel is planted beforehand. Within each class, we select the run with lowest validation cumulative regret before inspecting its kernel.

For each selected model, we fit kernel matrices κ_u to its predictions on fresh sequences of 501 observations:

$$g_{501} \approx x_{502}^\top \sum_{u=0}^{500} \kappa_u x_{501-u} y_{501-u}.$$

The fitted kernel aggregates evidence atoms into a belief, which is then applied to the final regressor. Writing $\xi = U^\top x$, we regress on $\xi_{502,i}\xi_{501-u,j}y_{501-u}$ for $i, j \in \{1, 2\}$; the coefficients, in original units, are $(U^\top \kappa_u U)_{ij}$. We plot the diagonal entries $\kappa_{u,i} := (U^\top \kappa_u U)_{ii}$: off-diagonal entries account for only 0.91% and 1.22% of the fitted attention and SSM kernels' squared Frobenius norms over $u = 1, \ldots, 500$.

We fit on 8192 episodes, rescaling each product to unit empirical root-mean-square without centering and minimizing mean squared error with ridge penalty 10^{-2}. To assess coefficient uncertainty, we use 100 random-sign (Rademacher) perturbations of the regression residuals, holding the trained models and probe inputs fixed. The shaded bands in Figure 12 show the central 95% of the resulting estimates separately for each coefficient.

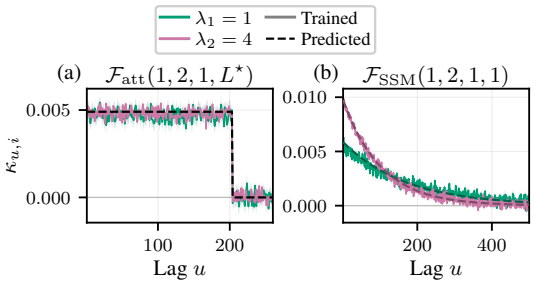


Figure 12: Optimal-kernel validation at $T = 4096$. *Trained*: kernels fitted to model predictions; *Predicted*: kernels from Prop. 1. Shading shows pointwise 95% bands over 100 residual perturbations.

Figure 12 shows that kernels fitted to the models' predictions closely follow the leading-order optima in Proposition 1: a uniform attention window and mode-dependent exponential decay for the SSM. Over $u = 1, \ldots, 500$, their relative Frobenius errors against these theoretical kernels are 13.5% and 15.6%, respectively. Both kernel-only fits explain 99.92% of the variation in the models' predictions on 2048 independent probe episodes ($R^2 = 0.9992$), showing that this kernel-form description accurately captures the selected models' prediction behavior.

D.3.2 Case 2.a: positional routing

Task and score. To test how depth and positional reach affect evidence assembly, we train the analytical models on scalar Gaussian regression with $p = 1$, delay $K = 4$, and horizon $T = 512$. Each episode draws independent standard normal θ, regressors x_t, and noises ϵ_t, with $y_t = \theta x_{t-K} + \epsilon_t$ for $t > K$ and $y_1 = \cdots = y_K = 0$. The model receives $z_t = (x_{t+1}, y_t, x_t)$ and predicts y_{t+1}. We report the Bayes-normalized performance score under squared loss:

$$S = 1 - \frac{\sum_{\text{episodes}} \sum_{t=1}^{T} (g_t - g_t^{\text{Bayes}})^2}{\sum_{\text{episodes}} \sum_{t=1}^{T} (g_t^{\text{Bayes}})^2}.$$

All T prediction times are included; zero prediction gives $S = 0$ and Bayes gives $S = 1$.

Models. Both linear attention and the fixed-transition SSM use $D \in \{1, 2\}$, one head, residual width 18, and $Hd \in \{1, 2, 4, 8, 16\}$. We sweep $L \in \{1, 2, 3, 4, 5, 8, 16, 32, 64, 96\}$ for attention and $d_{\text{conv}} \in \{1, 2, 3, 4, 5, 8\}$ for the SSM. All model parameters are trainable. Affine biases start at zero; SSM convolutions start as causal identities and input-independent transitions as $0.9I$. Attention's random query and key weights are scaled by $L^{-1/4}$ and output projections by $L^{-1/2}$; no routing lag is planted at initialization.

Training and evaluation. We train on noisy responses using squared loss; Bayes predictions are used only for evaluation. Each cell searches four paired initialization/data-stream seeds and learning rates $\{10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$. We use 12,000 constant-rate AdamW updates, fresh batches of 16 episodes, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01, and gradient clipping at 5. We evaluate checkpoints at 3000, 6000, and 12,000 updates and report the maximum over 60 scheduled candidates per cell. Selection and reporting use the same fixed set of 256 validation episodes.

Results. Figure 13 shows low scores for one-layer attention, whereas two layers benefit strongly from longer context. At depth two, $L = 2$ remains near zero, $L = 3, 4$ give partial gains, and $Hd \geq 2$ reaches about 0.92–0.94 at $L = 64, 96$. The one-layer SSM reaches about 0.94 at $d_{\text{conv}} \geq 5$; two layers already reach 0.95–0.99 at $d_{\text{conv}} = 3$ for $Hd \geq 2$. Thus additional depth enables successful learning with shorter convolutions. This is a benefit not observed in the practical Mamba-2 runs in Figure 4 (§D.1.3). The different task, training dynamics, or architectural choices such as selective transitions may account for this contrast.

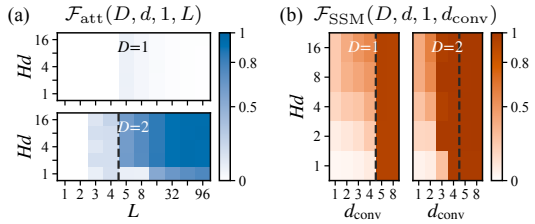


Figure 13: Positional routing in the analytically tractable classes ($p = 1$, $K = 4$, $T = 512$). Colors show the best validation score under the fixed search above. Dashed lines mark the single-layer reference $K + 1 = 5$.

These best scores describe successful configurations, not typical training outcomes. For example, at $d_{\text{conv}} = 3$ and $Hd \geq 2$, only 1–4 of the 20 seed–learning-rate runs per cell reach $S \geq 0.8$. The short-convolution depth advantage is therefore optimization-sensitive under this training protocol.

D.3.3 Case 2.b: content routing

Task and score. To test how nonlinear selection mechanisms of architectures affect content addressing, we use the Gaussian content-routing problem of §4.2 with $p = 1$, $R = 8$, $d_{\text{key}} = 6$, $T = 512$, and unit prior, regressor, and noise variances. The fixed codebook has maximum absolute pairwise inner product below $1/3$. Models receive the current and next keys along with (x_{t+1}, y_t, x_t); the routed regressor is not supplied. We use the Bayes-normalized squared-loss score S defined in §D.3.2, including all T prediction times.

Models. All models have depth two and residual width 18. Attention has two heads, $Hd \in \{2, 4, 8, 16\}$, and $L \in \{8, 16, 32, 64\}$; we compare all-linear attention with softmax in the first layer and linear attention in the second. The one-head SSM compares fixed transitions in both layers with a diagonal sigmoid transition in the first layer and a fixed transition in the second. Both retain ordinary bilinear writes and reads, with $d_{\text{conv}} \in \{1, 2, 4, 8\}$ and per-layer state widths $Hd \in \{1, 2, 4, 8, 16\}$. SSM convolutions start as causal identities, fixed transitions as $0.9I$, and sigmoid gates at the constant 0.9. Linear attention scales its random query and key weights by $L^{-1/4}$ and output projections by $L^{-1/2}$, as in Case 2.a; softmax attention keeps the default random initialization.

Training and evaluation. We train on noisy responses using squared loss and use the same 12,000-update AdamW protocol as §D.3.2: four seeds, five learning rates, and three checkpoints per run. Each cell reports the maximum over 60 scheduled candidates, with selection and reporting on the same fixed set of 256 validation episodes.

Results. In Figure 14, softmax attention reaches 0.83–0.93 for $Hd \geq 4$ and $L \geq 16$, whereas all-linear attention remains below 0.46. Neither SSM performs well at very small state widths. No displayed fixed-transition configuration reaches the success threshold $S \geq 0.8$; selective transitions enable success at $Hd = 16$, $d_{\text{conv}} = 8$, with score 0.820.

As stated in the previous section, best-run performance does not imply reliable optimization: the successful $Hd = 16$, $d_{\text{conv}} = 8$ cell reaches $S \geq 0.8$ in only one of its 20 seed–learning-rate runs.

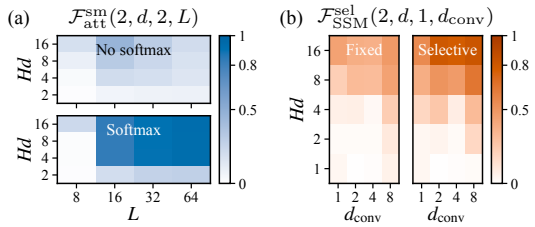


Figure 14: Content routing in the analytically tractable classes ($p = 1$, $R = 8$, $T = 512$, $D = 2$). Colors show the best validation score under the fixed search.

E CASE 1: FULL STATEMENTS AND PROOFS

For an overview of the notation and proof dependencies, see the notation guide and proof guide at the start of the appendix.

E.1 THEOREM 1 AND PROPOSITION 1: STATIONARY KERNELS

This section proves Theorem 1 and Proposition 1. We first introduce whitened coordinates and derive the regret of a stationary kernel (§E.1.1), then optimize over all such kernels (§E.1.2). Using the belief geometry in Figure 1 (§E.1.3), we connect the stationary-kernel analysis to attention (§E.1.4) and SSMs (§E.1.5). Section E.1.6 combines these results to complete the proof of Proposition 1.

E.1.1 WHITENED COORDINATES AND STATIONARY-KERNEL REGRET

We first express prediction error as belief-estimation error. Lemma 1 then gives the exact loss and regret of a stationary kernel, providing the objective we optimize in §E.1.2. Fix $p \in \mathbb{N}$, $\Sigma_x \succ 0$, $\Sigma_\theta \succ 0$, and $\sigma^2 > 0$, independently of the horizon, and consider

$$\theta \sim \mathcal{N}(0, \Sigma_\theta), \qquad x_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_x), \qquad y_t = x_t^\top \theta + \epsilon_t, \qquad \epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

with θ, $(x_t)_{t\geq 1}$, and $(\epsilon_t)_{t\geq 1}$ mutually independent. The token is $z_t = (x_{t+1}, y_t, x_t)$, and the predictor estimates the scalar y_{t+1}. The sufficient statistics

$$A_t := \sum_{s\leq t} x_s y_s, \qquad B_t := \sum_{s\leq t} x_s x_s^\top$$

give the posterior mean:

$$\hat{\theta}_t^{\text{Bayes}} = \left(\Sigma_\theta^{-1} + \sigma^{-2} B_t\right)^{-1} \sigma^{-2} A_t.$$

Introduce the whitened coordinates

$$\widetilde{x}_t := \Sigma_x^{-1/2} x_t, \qquad \widetilde{\theta} := \Sigma_x^{1/2}\theta, \qquad \widetilde{\Sigma}_\theta := \Sigma_x^{1/2}\Sigma_\theta\Sigma_x^{1/2} = U\,\mathrm{diag}(\lambda_1, \ldots, \lambda_p)U^\top. \tag{3}$$

Then $\widetilde{x}_t \sim \mathcal{N}(0, I_p)$, $\widetilde{\theta} \sim \mathcal{N}(0, \widetilde{\Sigma}_\theta)$, and $y_t = \widetilde{x}_t^\top \widetilde{\theta} + \epsilon_t$. The eigenvalues of $\widetilde{\Sigma}_\theta$ are the generalized belief variances in the fresh-regressor prediction geometry. The evidence innovation $\widetilde{x}_s y_s - \widetilde{\theta}$ has covariance Γ, with eigenvalues v_i:

$$\Gamma := \widetilde{\Sigma}_\theta + \left[\mathrm{tr}(\widetilde{\Sigma}_\theta) + \sigma^2\right] I_p, \qquad v_i := \lambda_i + \mathrm{tr}(\widetilde{\Sigma}_\theta) + \sigma^2. \tag{4}$$

We first record the fresh-regressor projection used throughout the section. Let $\mathcal{G}_t^- := \sigma(\widetilde{x}_{1:t}, y_{1:t})$. For an arbitrary scalar predictor g, define

$$\widetilde{\theta}_t(g) := \mathbb{E}[\widetilde{x}_{t+1} g_t \mid \mathcal{G}_t^-], \qquad g_t^\perp := g_t - \widetilde{x}_{t+1}^\top \widetilde{\theta}_t(g). \tag{5}$$

Since $\widetilde{x}_{t+1}$ is independent of $\mathcal{G}_t^-$ and has covariance I_p, $g_t^\perp$ is orthogonal to every $\widetilde{x}_{t+1}^\top a_t$ with $a_t \in L^2(\mathcal{G}_t^-; \mathbb{R}^p)$. The whitened posterior mean is

$$\widetilde{\theta}_t^{\text{Bayes}} = \left(\widetilde{\Sigma}_\theta^{-1} + \sigma^{-2}\sum_{s\leq t} \widetilde{x}_s \widetilde{x}_s^\top\right)^{-1} \sigma^{-2} \sum_{s\leq t} \widetilde{x}_s y_s.$$

Thus fresh-regressor projection and posterior orthogonality give

$$\ell_t(g) - \ell_t(g^{\text{Bayes}}) = \mathbb{E}\|\widetilde{\theta}_t(g) - \widetilde{\theta}_t^{\text{Bayes}}\|_2^2 + \mathbb{E}[(g_t^\perp)^2], \tag{6}$$

$$= \mathbb{E}\|\widetilde{\theta}_t(g) - \widetilde{\theta}\|_2^2 - e_t^{\text{Bayes}} + \mathbb{E}[(g_t^\perp)^2], \tag{7}$$

where the expected squared error of the Bayes belief and its cumulative sum are

$$e_t^{\text{Bayes}} := \mathbb{E}\,\mathrm{tr}\left[\left(\widetilde{\Sigma}_\theta^{-1} + \sigma^{-2}\sum_{s\leq t} \widetilde{x}_s \widetilde{x}_s^\top\right)^{-1}\right], \qquad \mathcal{E}_T^{\text{Bayes}} := \sum_{t\leq T} e_t^{\text{Bayes}}. \tag{8}$$

We suppress their dependence on the fixed distributional parameters. Equivalently, in the original coordinates,

$$\mathbb{E}\Big[\big(x_{t+1}^\top(\hat\theta_t-\theta)\big)^2\,\Big|\,\hat\theta_t,\theta\Big]=(\hat\theta_t-\theta)^\top\Sigma_x(\hat\theta_t-\theta)=\|\widetilde\theta_t-\widetilde\theta\|_2^2.$$

We first optimize over stationary-kernel predictors g of the form

$$g_t=x_{t+1}^\top\hat\theta_t(g),\qquad \hat\theta_t(g)=\sum_{u<t}\kappa_u x_{t-u}y_{t-u},$$

where $\kappa=(\kappa_u)_{u\ge0}$ is any deterministic matrix kernel. Under (3),

$$\widetilde\theta_t(g)=\Sigma_x^{1/2}\hat\theta_t(g)=\sum_{u<t}K_u\widetilde x_{t-u}y_{t-u},\quad K_u:=\Sigma_x^{1/2}\kappa_u\Sigma_x^{1/2},\quad M_t:=\sum_{u<t}K_u. \tag{9}$$

Here M_t is the kernel mass. For this predictor, we also write $\mathrm{Reg}_T(\kappa)=\mathrm{Reg}_T(K):=\mathrm{Reg}_T(g)$. Whitening leaves predictions and regret unchanged. Since the transformation is invertible, optimizing over the transformed stationary kernels is equivalent to optimizing over the original ones.

The next lemma makes the stationary kernel's bias–variance trade-off explicit, giving the regret formula we optimize to prove Theorem 1.

Lemma 1 (Exact loss and regret of a stationary kernel)**.** *For every deterministic stationary kernel $K=(K_u)_{u\ge0}$ in whitened coordinates, the corresponding predictor g satisfies*

$$\mathbb{E}\|\widetilde\theta_t(g)-\widetilde\theta\|_2^2=\mathrm{tr}\Big[(M_t-I_p)\widetilde\Sigma_\theta(M_t-I_p)^\top\Big]+\sum_{u<t}\mathrm{tr}\big[K_u\Gamma K_u^\top\big]\,, \tag{10}$$

$$\mathrm{Reg}_T(K)=\sum_{t\le T}\Big\{\mathrm{tr}\Big[(M_t-I_p)\widetilde\Sigma_\theta(M_t-I_p)^\top\Big]+\sum_{u<t}\mathrm{tr}\big[K_u\Gamma K_u^\top\big]\Big\}-\mathcal{E}_T^{\mathrm{Bayes}}. \tag{11}$$

The two terms in (10) are squared bias and variance, averaged over the prior: the bias comes from M_t-I_p, and the variance from the aggregated evidence. The loss is $\ell_t(g)=\mathbb{E}\|\widetilde\theta_t(g)-\widetilde\theta\|_2^2+\sigma^2$; the noise term cancels in regret.

Proof. Write

$$\widetilde x_s y_s=\widetilde\theta+\xi_s,\qquad \xi_s:=(\widetilde x_s\widetilde x_s^\top-I_p)\widetilde\theta+\widetilde x_s\epsilon_s.$$

Conditionally on $\widetilde\theta$, the innovations ξ_s are centered and independent across s. Hence

$$\mathbb{E}[\widetilde\theta\xi_s^\top]=0,\qquad \mathbb{E}[\xi_s\xi_r^\top]=0\quad(s\ne r).$$

For a standard Gaussian vector $\widetilde x$, Isserlis' identity gives

$$\mathbb{E}[\widetilde x\widetilde x^\top\widetilde\Sigma_\theta\widetilde x\widetilde x^\top]=2\widetilde\Sigma_\theta+\mathrm{tr}(\widetilde\Sigma_\theta)I_p.$$

Therefore

$$\begin{aligned}\mathbb{E}[\xi_s\xi_s^\top]&=\mathbb{E}[(\widetilde x\widetilde x^\top-I_p)\widetilde\Sigma_\theta(\widetilde x\widetilde x^\top-I_p)]+\sigma^2I_p\\&=\widetilde\Sigma_\theta+\big[\mathrm{tr}(\widetilde\Sigma_\theta)+\sigma^2\big]I_p=\Gamma.\end{aligned}$$

Since

$$\widetilde\theta_t(g)-\widetilde\theta=(M_t-I_p)\widetilde\theta+\sum_{u<t}K_u\xi_{t-u},$$

all bias–innovation and cross-time innovation terms vanish. Taking the squared norm proves (10). A stationary-kernel predictor has $g_t^\perp=0$, so (7) gives (11) after summing over time. □

Two direct checks are useful. The zero kernel has mean-squared belief-estimation error $\mathrm{tr}(\widetilde\Sigma_\theta)$ at every time. With one observation and an arbitrary matrix K,

$$\mathbb{E}\|K\widetilde x_1y_1-\widetilde\theta\|_2^2=2\,\mathrm{tr}(K\widetilde\Sigma_\theta K^\top)+\big[\mathrm{tr}(\widetilde\Sigma_\theta)+\sigma^2\big]\|K\|_F^2-2\,\mathrm{tr}(K\widetilde\Sigma_\theta)+\mathrm{tr}(\widetilde\Sigma_\theta),$$

which is the direct fourth-moment expansion of (10). When $p=1$, the innovation coefficient is $v_1=2\lambda_1+\sigma^2$.

E.1.2 Proof of Theorem 1

Starting from the exact regret formula in Lemma 1, Lemma 2 reduces stationary-kernel optimization to separate eigendirections, and Lemma 3 gives the sharp bound in each direction. Proposition 5 combines these results to prove Theorem 1 and compare exponential and uniform kernels.

Lemma 2 (Reduction to eigendirections)**.** *Let $\widetilde{K}_u := U^\top K_u U$ and $\widetilde{M}_t := \sum_{u<t} \widetilde{K}_u$. The sum of the bias and variance terms in (11) is exactly*

$$\sum_{r=1}^{p}\sum_{j=1}^{p}\left\{\lambda_j \sum_{t\le T}\big(\widetilde{M}_{t,rj} - \mathbf{1}\{r=j\}\big)^2 + v_j \sum_{t\le T}\sum_{u<t}\widetilde{K}^2_{u,rj}\right\}. \tag{12}$$

Every off-diagonal profile is therefore an independent nonnegative addition. An optimizer over all real stationary matrix kernels may be taken in the form

$$K_u = U\,\mathrm{diag}(\kappa_{u,1},\dots,\kappa_{u,p})U^\top,$$

and direction i has bias coefficient λ_i and innovation coefficient v_i.

Proof. Orthogonal invariance of the trace and $U^\top \widetilde{\Sigma}_\theta U = \mathrm{diag}(\lambda_1,\dots,\lambda_p)$ give

$$\mathrm{tr}[(M_t - I_p)\widetilde{\Sigma}_\theta (M_t - I_p)^\top] = \sum_{r,j}\lambda_j\big(\widetilde{M}_{t,rj} - \mathbf{1}\{r=j\}\big)^2.$$

Likewise, $U^\top \Gamma U = \mathrm{diag}(v_1,\dots,v_p)$ gives the second term of (12). If $r \neq j$, the corresponding profile has target zero in both positive quadratic terms, so setting it identically to zero cannot increase the cost. The remaining diagonal profiles are independent scalar problems. □

Each eigendirection gives a scalar stationary-kernel optimization problem. The following bound allows horizon-dependent bias and variance coefficients and identifies the optimal averaging scale.

Lemma 3 (Sharp scalar stationary bound)**.** *Let $\lambda_T, v_T > 0$ and define*

$$\omega_T := \sqrt{\frac{v_T T}{\lambda_T}}.$$

Suppose $\omega_T \to \infty$ and $\omega_T = o(T)$. These conditions hold automatically for fixed positive bias and variance coefficients, as in Theorem 1. For an arbitrary real lag profile $\kappa_0,\dots,\kappa_{T-1}$, let $m_t := \sum_{u<t}\kappa_u$. Then

$$\inf_{\kappa_0,\dots,\kappa_{T-1}\in\mathbb{R}}\left\{\lambda_T\sum_{t\le T}(m_t-1)^2 + v_T\sum_{t\le T}\sum_{u<t}\kappa_u^2\right\} = \sqrt{\lambda_T v_T T}\,(1+o(1)). \tag{13}$$

The lower bound is uniform over the lag profile, including profiles chosen as a function of T.

Proof. Set $a_0 := 1$ and $a_j := 1 - \sum_{u<j}\kappa_u$ for $1 \le j \le T$, so $\kappa_{j-1} = a_{j-1} - a_j$. The objective in (13) becomes

$$J_T(a) = \lambda_T\sum_{j=1}^{T} a_j^2 + v_T\sum_{j=1}^{T}(T-j+1)(a_{j-1}-a_j)^2. \tag{14}$$

Choose an integer m_T such that $\omega_T \ll m_T \ll T$; for example, $m_T = \lfloor\sqrt{T\omega_T}\rfloor$. Dropping the terms with $j > m_T$ and using $T - j + 1 \ge T - m_T + 1$ gives a constant-coefficient quadratic chain with edge coefficient $c_T := v_T(T - m_T + 1)$. If q_j is the minimum of its length-j tail as a multiple of a_0^2, dynamic programming gives

$$q_j = \frac{c_T(\lambda_T + q_{j-1})}{c_T + \lambda_T + q_{j-1}}, \qquad q_0 = 0.$$

The iteration increases to the positive fixed point

$$q_+ = \frac{\sqrt{\lambda_T^2 + 4\lambda_T c_T} - \lambda_T}{2} = \sqrt{\lambda_T v_T T}\,(1+o(1)).$$

Let

$$q_- := \frac{-\sqrt{\lambda_T^2 + 4\lambda_T c_T} - \lambda_T}{2} < 0$$

be the other fixed point. The exact cross-ratio identity for this Möbius iteration is

$$\frac{q_j - q_+}{q_j - q_-} = \rho_T^j \frac{q_0 - q_+}{q_0 - q_-}, \qquad \rho_T := \frac{c_T^2}{(c_T + \lambda_T + q_+)^2}.$$

Because $c_T/\lambda_T = \omega_T^2(1 + o(1))$, we have $1 - \rho_T = \Theta(1/\omega_T)$ and $q_+/|q_-| = 1 + o(1)$. Since $0 \le q_j \le q_+$, the cross-ratio identity therefore gives

$$0 \le 1 - \frac{q_j}{q_+} \le C\rho_T^j.$$

The condition $m_T/\omega_T \to \infty$ yields $q_{m_T} = q_+(1 - o(1))$, proving the lower bound.

For the matching upper bound, take the normalized exponential profile $\kappa_u = (1 - \alpha_T)\alpha_T^u$ with effective width $W_T := (1 + \alpha_T)/(1 - \alpha_T) = 2\omega_T$. Writing $D_T(\alpha) := \sum_{t=1}^T \alpha^{2t}$, direct geometric summation gives

$$J_T(a) = \lambda_T D_T(\alpha_T) + v_T \left(\frac{T}{W_T} - \frac{D_T(\alpha_T)}{W_T} \right).$$

Since $W_T \to \infty$ and $W_T = o(T)$,

$$D_T(\alpha_T) = \left(\frac{W_T}{4} - \frac{1}{2} + \frac{1}{4W_T} \right)(1 - \alpha_T^{2T}) = \frac{W_T}{4}(1 + o(1)).$$

It follows that

$$J_T(a) = \lambda_T \frac{W_T}{4}(1 + o(1)) + v_T \frac{T}{W_T}(1 + o(1)) = \sqrt{\lambda_T v_T T}\,(1 + o(1)).$$

□

Define the total bias and variance coefficients Λ and V and use the nearest-integer version of $L^\star(T)$ in the proofs:

$$\Lambda := \sum_{i=1}^p \lambda_i = \operatorname{tr}(\widetilde{\Sigma}_\theta), \qquad V := \sum_{i=1}^p v_i = (p+1)\operatorname{tr}(\widetilde{\Sigma}_\theta) + p\sigma^2, \tag{15}$$

$$\zeta := \sum_{i=1}^p \sqrt{\lambda_i v_i}, \qquad \bar{\zeta} := \sqrt{\Lambda V}, \tag{16}$$

$$L^\star(T) := \left\lfloor \sqrt{\frac{3VT}{\Lambda}} + \frac{1}{2} \right\rfloor. \tag{17}$$

Proposition 5 (Stationary floor)**.** *For fixed p, $\Sigma_x \succ 0$, $\Sigma_\theta \succ 0$, and $\sigma^2 > 0$, the infimum over deterministic real stationary matrix kernels K satisfies*

$$\inf_K \operatorname{Reg}_T(K) = \zeta\sqrt{T} + o(\sqrt{T}). \tag{18}$$

The modewise normalized exponential

$$K_u^{\exp} = U \operatorname{diag}((1 - \alpha_i)\alpha_i^u)_{i=1}^p U^\top, \qquad \frac{1 + \alpha_i}{1 - \alpha_i} = 2\sqrt{\frac{v_i T}{\lambda_i}}, \tag{19}$$

attains (18) to leading order.

For comparison, the identity-scaled uniform kernel

$$K_u^{\mathrm{unif}} := \frac{1}{L^\star(T)} I_p \mathbf{1}\{0 \le u < L^\star(T)\}$$

has

$$\operatorname{Reg}_T(K^{\mathrm{unif}}) = \frac{2}{\sqrt{3}}\bar{\zeta}\sqrt{T} + o(\sqrt{T}). \tag{20}$$

The kernel parameters may depend on the horizon and the fixed distributional parameters, but one parameter setting is held fixed throughout the stream.

Proof. Lemma 2 reduces the optimization to p scalar problems. Applying Lemma 3 with $(\lambda_T, v_T) = (\lambda_i, v_i)$ gives

$$\inf_K \{\mathrm{Reg}_T(K) + \mathcal{E}_T^{\mathrm{Bayes}}\} = \sqrt{T}\sum_{i=1}^{p}\sqrt{\lambda_i v_i} + o(\sqrt{T}),$$

where fixed p permits summing the directionwise remainders. The exponential profile used in the upper bound of that lemma in each eigendirection is exactly (19).

It remains to control the Bayes subtraction. Let $\widetilde{B}_t := \Sigma_x^{-1/2} B_t \Sigma_x^{-1/2} = \sum_{s\le t} \widetilde{x}_s \widetilde{x}_s^\top$ and $D_0 := \sigma^2 \widetilde{\Sigma}_\theta^{-1}$. Then

$$e_t^{\mathrm{Bayes}} = \sigma^2 \mathbb{E}\,\mathrm{tr}(\widetilde{B}_t + D_0)^{-1}.$$

For all sufficiently large t, $\widetilde{B}_t$ is invertible and

$$0 \preceq \widetilde{B}_t^{-1} - (\widetilde{B}_t + D_0)^{-1} \preceq \widetilde{B}_t^{-1} D_0 \widetilde{B}_t^{-1} \preceq \|D_0\|_{\mathrm{op}} \widetilde{B}_t^{-2}.$$

The fixed-dimensional inverse-Wishart identities give

$$\mathbb{E}\,\mathrm{tr}(\widetilde{B}_t^{-1}) = \frac{p}{t-p-1}, \qquad \mathbb{E}\,\mathrm{tr}(\widetilde{B}_t^{-2}) = O(t^{-2}), \qquad p \text{ fixed}.$$

The finitely many earlier terms are bounded by $\mathrm{tr}(\widetilde{\Sigma}_\theta)$, so

$$\mathcal{E}_T^{\mathrm{Bayes}} = p\sigma^2 \log T + O(1) = o(\sqrt{T}). \tag{21}$$

Subtracting this term proves (18).

For the uniform kernel, summing its cold-start bias and variance terms gives exactly

$$\Lambda\left(\frac{L}{3} - \frac{1}{2} + \frac{1}{6L}\right) + V\left(\frac{T}{L} - \frac{1}{2} + \frac{1}{2L}\right).$$

The nearest-integer choice (17) minimizes its leading terms, yielding $\frac{2}{\sqrt{3}}\bar{\zeta}\sqrt{T} + o(\sqrt{T})$. Equation (21) proves (20). □

E.1.3 KERNEL-FORM PROJECTION

We now connect the kernel analysis to the architecture classes through belief geometry (Figure 1). Lemma 4 shows that, for degree-two beliefs, projecting onto kernel form with deterministic kernel weights can only decrease prediction error. This allows the attention and SSM proofs below to focus on the stationary kernels their architectures realize.

Let $\mathcal{Q}_{2,t}$ be the space of $\mathbb{R}^p$-valued polynomials of total degree at most two in $(\widetilde{x}_{1:t}, y_{1:t})$, and let

$$\mathcal{S}_t := \left\{ \sum_{s\le t} K_s \widetilde{x}_s y_s : K_s \in \mathbb{R}^{p\times p} \right\}.$$

Here the matrices are deterministic. Thus $\mathcal{S}_t$ is contained in the whitened kernel-form space $\Sigma_x^{1/2}\mathcal{K}_t$, which also allows regressor-dependent kernel weights. This projection is a further reduction, distinct from Π_t: its remainder need not lie outside kernel form.

Lemma 4 (Projection onto deterministic kernel form)**.** *Every $f \in \mathcal{Q}_{2,t}$ has an orthogonal decomposition $f = d + n$ with $d \in \mathcal{S}_t$ and $n \perp \mathcal{S}_t$. Moreover, $\mathbb{E}[n^\top \widetilde{\theta}] = 0$. Consequently,*

$$\mathbb{E}\|f - \widetilde{\theta}\|_2^2 = \mathbb{E}\|d - \widetilde{\theta}\|_2^2 + \mathbb{E}\|n\|_2^2. \tag{22}$$

Proof. By linearity, it suffices to check scalar monomials of degree at most two. Constants, design-only monomials, and terms containing two response factors have zero inner product with $\widetilde{\theta}$ by centering and the joint sign symmetry $(\widetilde{\theta}, \epsilon_{1:t}) \mapsto -(\widetilde{\theta}, \epsilon_{1:t})$. Linear response terms also have zero inner product with $\widetilde{\theta}$, because $\mathbb{E}[\widetilde{\theta} y_s] = \widetilde{\Sigma}_\theta \mathbb{E}[\widetilde{x}_s] = 0$. Finally,

$$\mathbb{E}[\widetilde{\theta}\widetilde{x}_{r,i} y_s] = \widetilde{\Sigma}_\theta e_i\, \mathbf{1}\{r = s\}. \tag{23}$$

Since $\widetilde{\Sigma}_\theta \succ 0$, the only degree-two monomials correlated with $\widetilde{\theta}$ are the coordinates of the evidence atoms $\widetilde{x}_s y_s$.

The same sign symmetry and design parity show that every remaining monomial is orthogonal to every evidence atom. Terms with zero or two response factors follow from the joint response-sign symmetry after multiplication by $\widetilde{x}_k y_k$. For a linear response term, $\mathbb{E}[y_r \widetilde{x}_k y_k] = 0$: when $r = k$, reflect $(\widetilde{x}_k, \epsilon_k) \mapsto (-\widetilde{x}_k, -\epsilon_k)$; when $r \neq k$, conditioning on the designs leaves $\mathbb{E}[y_r y_k \mid \widetilde{x}] = \widetilde{x}_r^\top \widetilde{\Sigma}_\theta \widetilde{x}_k$, whose remaining expectation vanishes by design parity. For a cross-atom $\widetilde{x}_{r,i} y_s$ with $r \neq s$, conditioning on the designs again reduces its product with $\widetilde{x}_k y_k$ to design monomials. If $k \in \{r, s\}$, the other independent design vector has odd degree; if $k \notin \{r, s\}$, $\widetilde{x}_r$ remains an independent centered factor. Orthogonal projection onto $\mathcal{S}_t$ therefore leaves a remainder orthogonal both to $\mathcal{S}_t$ and to $\widetilde{\theta}$, proving (22). □

E.1.4 One-layer attention

We next characterize one-layer attention through the aggregation kernels it realizes. Lemma 5 shows that the aggregation kernel assigns the same matrix weight, of rank at most Hd, to every past evidence atom within the context window. Lemma 6 then shows that, at sufficient width, uniform kernels achieve the leading-order optimum within the one-layer attention class at every context length.

Lemma 5 (One-layer attention projection and kernel rank)**.** *Fix $d_e, d, H, L \geq 1$. For every $g \in \mathcal{F}_{\mathrm{att}}(1, d, H, L)$, there are deterministic matrices $K, K_0 \in \mathbb{R}^{p \times p}$ such that*

$$\widetilde{\theta}_t(g) = K \sum_{s=\max(1,t-L+1)}^{t-1} \widetilde{x}_s y_s + K_0 \widetilde{x}_t y_t + \widetilde{\theta}_t^\perp, \qquad \widetilde{\theta}_t^\perp \perp \mathcal{S}_t, \quad \mathbb{E}[(\widetilde{\theta}_t^\perp)^\top \widetilde{\theta}] = 0. \tag{24}$$

Here $\widetilde{\theta}_t^\perp$ is the orthogonal remainder after projecting $\widetilde{\theta}_t(g)$ onto $\mathcal{S}_t$. Moreover, the common past-lag kernel weight admits a headwise factorization

$$K = \sum_{b=1}^{H} A_b C_b, \qquad A_b \in \mathbb{R}^{p \times d}, \quad C_b \in \mathbb{R}^{d \times p}, \tag{25}$$

and hence $\mathrm{rank}(K) \leq Hd$. When $d_e \geq 2p+1$ and $Hd \geq p$, a distributed-head construction realizes $K = K_0 = U \,\mathrm{diag}(a_1, \ldots, a_p) U^\top$ for any real $a_1, \ldots, a_p$, with $\widetilde{\theta}_t^\perp = g_t^\perp = 0$.

Proof. Whitening is an invertible linear change to the two regressor blocks of each token, so it can be absorbed into the initial embedding and the affine query, key, and value maps without changing (d, H, L). For the lower-bound projection, homogenize the whitened token expressions as linear forms in $\bar{z}_s := (1, \widetilde{x}_{s+1}, y_s, \widetilde{x}_s)$. This represents the legal affine maps through a constant coordinate. Let $J_+, J_0 \in \mathbb{R}^{(2p+2) \times p}$ and $e_y \in \mathbb{R}^{2p+2}$ select $\widetilde{x}_{s+1}$, $\widetilde{x}_s$, and y_s, respectively, from $\bar{z}_s$. After composing the embedding with the affine query and key maps, and combining the value, output, and final prediction maps into a scalar map, head b contributes

$$a_b^\top \bar{z}_t + \sum_{s=\max(1,t-L+1)}^{t} \left[p_{t-s}^{(b)} + (\mathsf{Q}_b \bar{z}_t)^\top (\mathsf{K}_b \bar{z}_s) \right] (v_b^\top \bar{z}_s),$$

where $\mathsf{Q}_b, \mathsf{K}_b \in \mathbb{R}^{d \times (2p+2)}$ and $v_b \in \mathbb{R}^{2p+2}$ are deterministic. For $s < t$, multiplication by $\widetilde{x}_{t+1}$ and conditional expectation over the fresh regressor give

$$A_b (\mathsf{K}_b \bar{z}_s)(v_b^\top \bar{z}_s), \qquad A_b := (\mathsf{Q}_b J_+)^\top \in \mathbb{R}^{p \times d}.$$

The positional-bias term has zero fresh-regressor projection. Define

$$C_b := (e_y^\top v_b) \mathsf{K}_b J_0 + (\mathsf{K}_b e_y) v_b^\top J_0 \in \mathbb{R}^{d \times p}.$$

The two summands place $\widetilde{x}_s$ in the key and y_s in the value, or vice versa. Thus the kernel weight on the same-time evidence atom $\widetilde{x}_s y_s$ from head b is $A_b C_b$. All remaining source-token monomials are placed in $\widetilde{\theta}_t^\perp$ by Lemma 4. Because the maps are shared over positions, this weight is independent of s and of the lag. Summing heads gives (25) and $\mathrm{rank}(K) \leq Hd$.

At $s = t$, the positional term contributes only a degree-zero nuisance. The bilinear score–value term is cubic in $\bar{z}_t$; after multiplication by the fresh Gaussian regressor $\widetilde{x}_{t+1}$, only terms with one or three internal fresh-regressor factors survive conditional expectation. The result has degree at most two in $(\widetilde{x}_t, y_t)$ and may supply a separate current-atom kernel weight K_0. Lemma 4 places every other resulting monomial in $\widetilde{\theta}_t^{\perp}$, which proves (24).

For the witness, use the first $2p + 1$ residual coordinates for the embedding

$$e_s^{(0)} = (\widetilde{x}_{s+1}, \widetilde{x}_s, y_s, 0_{d_e-(2p+1)}) \in \mathbb{R}^{d_e}.$$

Since $Hd \geq p$, partition $[p]$ into sets $I_1, \ldots, I_H$ with $|I_b| \leq d$. Head b uses the unnormalized score

$$\sum_{i \in I_b} a_i (u_i^\top \widetilde{x}_{t+1})(u_i^\top \widetilde{x}_s)$$

where u_i is column i of U. Its value is y_s, and all head-output maps write into the shared label coordinate. In one designated head choose the relative positional bias $p_0 = -1$ and $p_u = 0$ for $1 \leq u < L$; set every positional bias in the other heads to zero. The final map reads the label coordinate. Summing the heads recovers $\widetilde{x}_{t+1}^\top K \widetilde{x}_s$, while the designated lag-zero bias cancels the residual contribution y_t. Hence the output is exactly

$$g_t = y_t + \sum_{s=\max(1,t-L+1)}^{t} \left[\widetilde{x}_{t+1}^\top K \widetilde{x}_s - \mathbf{1}\{s = t\}\right] y_s = \widetilde{x}_{t+1}^\top K \sum_{s=\max(1,t-L+1)}^{t} \widetilde{x}_s y_s.$$

Each head uses at most d query/key coordinates, with no nuisance or fresh-regressor residual. Taking $a_i = 1/L$ recovers the normalized uniform kernel. □

Lemma 5 allows the current evidence atom to receive a different kernel weight from past atoms. The next lemma accounts for the additional error this freedom introduces and shows that uniform kernels remain optimal to leading order within this class at every context length.

Lemma 6 (One-layer attention leading-order optimal kernel)**.** *Fix the covariance parameters above and take $1 \leq L \leq T$; larger contexts give the same windows as $L = T$. Write $\tau_t := \min(t, L)$ and $N_j := \sum_{t \leq T} \tau_t^j$ for $j = 1, 2$. The best uniform kernel has*

$$\bar{K} := U \operatorname{diag}(a_1, \ldots, a_p) U^\top, \qquad a_i := \frac{\lambda_i N_1}{\lambda_i N_2 + v_i N_1}, \qquad K_u^{\mathrm{unif},\star} := \bar{K} \mathbf{1}\{0 \leq u < L\}, \quad (26)$$

and cumulative mean-squared belief-estimation error

$$\mathcal{E}_{L,T} := \sum_{i=1}^{p} \left(\lambda_i T - \frac{\lambda_i^2 N_1^2}{\lambda_i N_2 + v_i N_1} \right). \quad (27)$$

For every $d_e, d, H \geq 1$,

$$\mathrm{Reg}_T\big(\mathcal{F}_{\mathrm{att}}(1, d, H, L)\big) + \mathcal{E}_T^{\mathrm{Bayes}} \geq \mathcal{E}_{L,T} - O(L/T). \quad (28)$$

If $Hd \geq p$ and $d_e \geq 2p + 1$, this uniform kernel can be realized by an attention model in the class, giving the reverse bound with $\mathcal{E}_{L,T}$ in place of $\mathcal{E}_{L,T} - O(L/T)$. In the original coordinates its kernel weights within the context window are

$$\kappa_u^{\mathrm{unif},\star} = \Sigma_x^{-1/2} \bar{K} \Sigma_x^{-1/2} = \frac{1}{L} \Sigma_x^{-1} + O(L^{-2} + T^{-1}), \qquad 0 \leq u < L, \quad (29)$$

with the same matrix correction at every lag and zero weights otherwise. The bounds above are uniform in L, d_e, d, H at fixed covariance parameters; matrix remainders are measured in Frobenius norm. In particular, when $L = \Theta(\sqrt{T})$,

$$\mathcal{E}_{L,T} = \frac{\Lambda L}{3} + \frac{VT}{L} + O(1), \quad (30)$$

and the sufficient-width class matches this kernel's regret up to $O(1)$.

Proof. Set $\Delta := K_0 - K$. The two projection identities and Lemma 1 give the exact cost

$$\mathrm{Reg}_T(g) + \mathcal{E}_T^{\mathrm{Bayes}} = \sum_{t\le T}\Big[\|(\tau_t K + \Delta - I_p)\widetilde{\Sigma}_\theta^{1/2}\|_F^2 + (\tau_t - 1)\|K\Gamma^{1/2}\|_F^2 + \|(K+\Delta)\Gamma^{1/2}\|_F^2 + \mathbb{E}\|\widetilde{\theta}_t^{\perp}\|_2^2 + \mathbb{E}[(g_t^{\perp})^2]\Big]. \tag{31}$$

For equal kernel weights, Lemma 2 gives the directionwise quadratic $\lambda_i T - 2\lambda_i N_1 a_i + (\lambda_i N_2 + v_i N_1)a_i^2$. Its minimizer and value are (26)–(27), and Lemma 5 supplies the legal witness. When $L = 1$, only K_0 remains, so this is also the exact class optimum at sufficient width. Henceforth take $L \ge 2$.

Cost of changing the current-atom kernel weight. In the cubic expansion used in the attention projection proof, the common evidence contribution is $\widetilde{x}_{t+1}^\top K \widetilde{x}_s y_s$. The only other placements of its three factors that contribute to the current-atom kernel weight at $s = t$ are

$$y_t \widetilde{x}_{s+1}^\top D_1 \widetilde{x}_s, \qquad y_s \widetilde{x}_{s+1}^\top D_2 \widetilde{x}_t, \qquad D_1 + D_2 = \Delta,$$

where D_1, D_2 collect the coefficients across heads. Lower-degree terms cannot contribute to this kernel weight, and terms with three fresh factors contribute only a constant to the implicit belief, orthogonal to the evidence atoms.

For $s \in J_t$ with $s \le t-2$, let $R_{t,s}$ be the sum of these two terms. Under the sign symmetries $(\widetilde{x}_j, \epsilon_j) \mapsto (-\widetilde{x}_j, -\epsilon_j)$ it is odd exactly at the three sites $\{s, s+1, t\}$, and it is also odd under the joint response-sign symmetry used in Lemma 4. Only source s contains both earlier sites, so these are the only legal degree-three terms with both properties. Different s give orthogonal components of $g_t^{\perp}$: they contain no fresh regressor and their sign patterns differ. Affine terms, positional biases, and residuals have too few factors to cancel them. We omit $s = t-1$, where the sites overlap. The Gaussian moments used above give

$$\mathbb{E}R_{t,s}^2 = \tfrac{1}{2}\|\Delta\Gamma^{1/2}\|_F^2 + \tfrac{1}{2}\|(D_1 - D_2)((\Lambda+\sigma^2)I_p - \widetilde{\Sigma}_\theta)^{1/2}\|_F^2.$$

Since $(\Lambda+\sigma^2)I_p - \widetilde{\Sigma}_\theta \succ 0$, orthogonality yields

$$\sum_{t\le T}\mathbb{E}[(g_t^{\perp})^2] \ge \frac{m}{2}\|\Delta\Gamma^{1/2}\|_F^2, \qquad m := \sum_{t\le T}(\tau_t - 2)_+. \tag{32}$$

Comparison with the uniform kernel. Write $\widetilde{K} = U^\top K U$ and $\widetilde{\Delta} = U^\top \Delta U$. To complete the same quadratic, put

$$P_i := \lambda_i N_2 + v_i N_1, \qquad J_i := \lambda_i N_1 + v_i T,$$
$$f_i := \frac{\lambda_i^2(TN_2 - N_1^2)}{P_i}, \qquad h_i := T(\lambda_i + v_i) - \frac{J_i^2}{P_i} + \frac{m v_i}{2}.$$

Combining (31) and (32), then completing squares column by column, gives

$$\mathrm{Reg}_T(g) + \mathcal{E}_T^{\mathrm{Bayes}} \ge \mathcal{E}_{L,T} + \sum_{i=1}^{p}\Big[P_i\big\|(\widetilde{K} - a_i I_p + (J_i/P_i)\widetilde{\Delta})e_i\big\|_2^2 + h_i\|\widetilde{\Delta}e_i\|_2^2 - 2f_i\widetilde{\Delta}_{ii}\Big]. \tag{33}$$

Here e_i is the ith coordinate vector. Uniformly for $2 \le L \le T$, $P_i \asymp TL^2$, $J_i/P_i = O(L^{-1})$, and $0 \le f_i \le \lambda_i L$. Indeed, $TL/2 \le N_1 \le TL$, $N_2 \ge TL^2/3$, and $TN_2 - N_1^2 \le T\sum_t (L-\tau_t)^2 \le TL^3/3$. Minimizing only the innovation part in (31) also gives

$$h_i - \frac{m v_i}{2} \ge v_i T(1 - T/N_1) \ge 0.$$

Together with $m = (L-2)(T - (L+1)/2)$, this implies $h_i \ge v_i T(L-1)/12$. For $L = 2$, where $m = 0$, the innovation term alone gives $h_i \ge v_i T/3$. Completing the remaining square in $\widetilde{\Delta}e_i$ therefore lowers $\mathcal{E}_{L,T}$ by at most $\sum_i f_i^2/h_i = O(L/T)$, proving (28). The upper bound comes from the uniform witness, not from assuming that this relaxed minimizer is legal.

At sufficient width, the same squares show that any predictor within $o(L/T)$ of the class infimum satisfies

$$\|K - \bar{K}\|_F = O((T\sqrt{L})^{-1}), \qquad \|\Delta\|_F = O(T^{-1}). \tag{34}$$

To obtain the simpler kernel expression, use

$$a_i = \left(\frac{N_2}{N_1} + \frac{v_i}{\lambda_i}\right)^{-1}, \qquad 0 \le L - \frac{N_2}{N_1} = \frac{L(L^2-1)}{6N_1} \le \frac{L^2-1}{3T}, \qquad \frac{N_2}{N_1} \ge \frac{L}{3}.$$

These imply $a_i = 1/L + O(L^{-2} + T^{-1})$, and transforming back by $\Sigma_x^{-1/2}$ proves (29). When $L = \Theta(\sqrt{T})$, $\bar{K} - I_p/L = O(T^{-1})$ and the uniform-kernel quadratic has coefficient $P_i = \Theta(TL^2)$ on the squared amplitude. Optimizing its amplitudes therefore improves the normalized uniform cost computed in Proposition 5 by only $O(1)$. This proves (30). □

Full-context attention. The distinction between an optimal amplitude and the normalization $1/L$ matters at $L = T$. Here $N_1 = T(T+1)/2$, $N_2 = T(T+1)(2T+1)/6$, so

$$a_i = \frac{3\lambda_i}{\lambda_i(2T+1) + 3v_i} = \frac{3}{2T} + O(T^{-2}), \qquad \mathcal{E}_{T,T} = \frac{\Lambda T}{4} + O(1).$$

Thus, at $Hd \ge p$ and $d_e \ge 2p+1$,

$$\mathrm{Reg}_T\big(\mathcal{F}_{\mathrm{att}}(1,d,H,T)\big) = \frac{\Lambda T}{4} - p\sigma^2 \log T + O(1). \tag{35}$$

Moreover, every predictor within $o(1)$ of the class infimum has $K = 3I_p/(2T) + O(T^{-3/2})$ by (34). The current-atom kernel weight is also of order $1/T$ in every direction. To see its lower bound, retain the m-term in (33) and use $m/T^2 \to 1/2$, $f_i/T \to \lambda_i/4$, and $h_i - mv_i/2 \ge 0$. Comparison with the uniform witness gives

$$\|(T\Delta - \widetilde{\Sigma}_\theta \Gamma^{-1})\Gamma^{1/2}\|_F^2 \le \mathrm{tr}(\widetilde{\Sigma}_\theta^2 \Gamma^{-1}) + o(1),$$

so $\|T\Delta - \widetilde{\Sigma}_\theta\Gamma^{-1}\|_{\mathrm{op}} \le \Lambda/(\Lambda+\sigma^2) + o(1)$. Since $\Lambda/(\Lambda+\sigma^2) < 1$ and $0 \preceq \widetilde{\Sigma}_\theta\Gamma^{-1} \prec I_p/2$, eventually

$$\frac{1}{2T} I_p \preceq \frac{K_0 + K_0^\top}{2}, \qquad \|K_0\|_{\mathrm{op}} \le \frac{3}{T}. \tag{36}$$

Transforming back by $\Sigma_x^{-1/2}$ gives the $\Theta(1/T)$ evidence weights described in the main text, including the current atom. The kernel mass is $tK + \Delta = 3tI_p/(2T) + O(T^{-1/2})$, uniformly for $t \le T$. It remains bounded away from I_p throughout an initial constant fraction of the horizon, explaining the linear cold-start bias.

E.1.5 Fixed-transition SSMs

We next apply the same projection to fixed-transition SSMs. Lemma 7 shows that the aggregation kernels obtained after projection are stationary and constructs an exponential kernel that attains the floor in Theorem 1 at sufficient width.

Lemma 7 (Fixed-transition SSM projection)**.** *Fix $d_e, d, H \ge 1$. Concatenate the H heads of the one-layer, identity-convolution, fixed-transition SSM into a state with block-diagonal transition $\Phi \in \mathbb{R}^{Hd \times Hd}$. Every $g \in \mathcal{F}_{\mathrm{SSM}}(1,d,H,1)$ has*

$$\widetilde{\theta}_t(g) = \sum_{u<t} K_u \widetilde{x}_{t-u} y_{t-u} + \widetilde{\theta}_t^{\perp}, \qquad \widetilde{\theta}_t^{\perp} \perp \mathcal{S}_t, \quad \mathbb{E}[(\widetilde{\theta}_t^{\perp})^\top \widetilde{\theta}] = 0, \tag{37}$$

where

$$K_u = W_{\mathrm{read}} \Phi^u W_{\mathrm{write}} \quad (u \ge 1), \qquad K_0 \in \mathbb{R}^{p\times p} \text{ is separate}, \tag{38}$$

for deterministic effective read and evidence-write maps $W_{\mathrm{read}} \in \mathbb{R}^{p \times Hd}$ and $W_{\mathrm{write}} \in \mathbb{R}^{Hd \times p}$. Hence every class member obeys the global stationary floor. If $Hd \ge p$ and $d_e \ge 2p+2$, the class contains the modewise exponential witness (19).

Proof. Whitening the two regressor blocks is absorbed into the initial embedding and the SSM's affine write and read maps. After concatenating heads,

$$h_t = \Phi h_{t-1} + u_t = \sum_{s \le t} \Phi^{t-s} u_s.$$

For $s < t$, computing the implicit belief from the ordinary token–state read gives a fixed map W_{read} from the state. More explicitly, after the residual update and final prediction the output can be written

$$g_t = r_0^\top e_t^{(0)} + \widetilde{x}_{t+1}^\top W_{\mathrm{read}} h_t + \bar{q}_t^\top h_t,$$

where r_0 is deterministic and $\bar{q}_t$ is $\mathcal{G}_t^-$-measurable. The direct residual term contributes to the implicit belief only a component orthogonal to the evidence atoms. On the old-state part Φh_{t-1}, the last term, $\bar{q}_t^\top \Phi h_{t-1}$, has zero fresh-regressor projection. For a historical write at time $s < t$, the token field $\widetilde{x}_{s+1}$ belongs to $\mathcal{G}_t^-$ at the eventual read time. We may therefore apply Lemma 4 in $\mathcal{Q}_{2,t}$; the cross-atom parity calculation in its proof shows that terms involving $\widetilde{x}_{s+1}$ remain orthogonal to the evidence atoms. Applied componentwise, the lemma thus decomposes the write's bilinear part as $W_{\mathrm{write}}\widetilde{x}_s y_s$ plus channels orthogonal to the evidence atoms. Deterministic linear maps preserve this orthogonality, so the lag-u kernel weight is $W_{\mathrm{read}}\Phi^u W_{\mathrm{write}}$ for $u \geq 1$.

At $s = t$, the fresh regressor also enters the current bilinear write, and computing the implicit belief can produce a separate K_0. The current token contributes at most degree three to g_t; multiplying by the fresh vector and taking conditional expectation leaves a polynomial of degree at most two in $(\widetilde{x}_t, y_t)$, so Lemma 4 applies. All remaining projected channels form $\widetilde{\theta}_t^\perp$, and the component outside the fresh-regressor subspace forms $g_t^\perp$. Equations (22) and (7) make both nonnegative additions to the stationary-kernel cost.

For the witness, use the first $2p + 1$ residual coordinates for the embedding

$$e_s^{(0)} = (\widetilde{x}_{s+1}, \widetilde{x}_s, y_s, 0_{d_e-(2p+1)}) \in \mathbb{R}^{d_e}.$$

Distribute the p eigendirections among the H heads, with at most d directions per head. This is possible when $Hd \geq p$. For $a_{i,t} := u_i^\top \widetilde{x}_t$, assign one state coordinate satisfying

$$h_{i,t} = \alpha_i h_{i,t-1} + a_{i,t} y_t.$$

The bilinear write is legal in the identity-convolution class. Read that coordinate with $(1 - \alpha_i)a_{i,t+1}$, sum the head-local outputs into one clean residual coordinate, and let the final prediction read that coordinate:

$$g_t = \sum_{i=1}^{p} (1 - \alpha_i) a_{i,t+1} \sum_{u<t} \alpha_i^u a_{i,t-u} y_{t-u}.$$

Each head transition may contain distinct diagonal entries, and the modes do not communicate. The resulting kernel is exactly (19), with $\widetilde{\theta}_t^\perp = g_t^\perp = 0$. □

E.1.6 Proof of Proposition 1

We first use the two architecture projections to establish the width floors in Lemma 8. We then combine these floors with the kernel bounds and constructions to prove Proposition 1.

Lemma 8 (Width floors)**.** *Define*

$$d_i := \frac{\lambda_i v_i}{\lambda_i + v_i}, \tag{39}$$

and denote their increasing rearrangement by $d_{(1)} \leq \cdots \leq d_{(p)}$. For every $g \in \mathcal{F}_{\mathrm{att}}(1, d, H, L)$, let K be the common past-lag kernel weight in Lemma 5. Then

$$\mathrm{Reg}_T(g) \geq T \sum_{j=1}^{p-\mathrm{rank}(K)} d_{(j)} - \mathcal{E}_T^{\mathrm{Bayes}} \geq T \sum_{j=1}^{(p-Hd)_+} d_{(j)} - \mathcal{E}_T^{\mathrm{Bayes}}. \tag{40}$$

Similarly, every $g \in \mathcal{F}_{\mathrm{SSM}}(1, d, H, 1)$ satisfies

$$\mathrm{Reg}_T(g) \geq T \sum_{j=1}^{(p-Hd)_+} d_{(j)} - \mathcal{E}_T^{\mathrm{Bayes}}. \tag{41}$$

Consequently, at fixed $p, \widetilde{\Sigma}_\theta, \sigma^2$, both classes incur $\Omega(T)$ regret whenever $Hd < p$. This holds uniformly over d_e and, for attention, L, including horizon-dependent choices.

Proof. For attention, let P be the orthogonal projector onto $\mathrm{col}(K)^\perp$ and put

$$a_t := K \sum_{s=\max(1,t-L+1)}^{t-1} \widetilde{x}_s y_s + K_0 \widetilde{x}_t y_t.$$

Since $a_t \in \mathcal{S}_t$, the projection lemma gives

$$\mathbb{E}\|\widetilde{\theta}_t(g) - \widetilde{\theta}\|_2^2 = \mathbb{E}\|a_t - \widetilde{\theta}\|_2^2 + \mathbb{E}\|\widetilde{\theta}_t^\perp\|_2^2 \geq \mathbb{E}\|P(a_t - \widetilde{\theta})\|_2^2.$$

Equation (7) likewise discards $g_t^\perp$ as a nonnegative contribution. Since $PK = 0$, keeping only the projected current-atom error gives, at every time,

$$\mathbb{E}\|\widetilde{\theta}_t(g) - \widetilde{\theta}\|_2^2 \geq \mathrm{tr}\Big[P(K_0 - I_p)\widetilde{\Sigma}_\theta(K_0 - I_p)^\top P\Big] + \mathrm{tr}\big[PK_0\Gamma K_0^\top P\big].$$

For the SSM, instead let P project onto $\mathrm{col}(W_{\mathrm{read}})^\perp$. Then $\mathrm{rank}(P) \geq (p - Hd)_+$ and $PK_u = 0$ for every $u \geq 1$. Lemma 1 and the SSM projection give the same displayed lower bound.

This SSM argument also allows independent write/read feature widths. Linear combinations of bilinear write features remain quadratic, while reads pairing affine token features with linear state projections reduce to the same token–state read. Thus the projection proof still gives (38) through an Hd-dimensional state.

Set $X := PK_0$, so $PX = X$. Minimizing the displayed quadratic over this relaxed variable gives

$$X^\star = P\widetilde{\Sigma}_\theta(\widetilde{\Sigma}_\theta + \Gamma)^{-1}$$

and minimum

$$\mathrm{tr}\Big[P\{\widetilde{\Sigma}_\theta - \widetilde{\Sigma}_\theta(\widetilde{\Sigma}_\theta + \Gamma)^{-1}\widetilde{\Sigma}_\theta\}\Big] = \mathrm{tr}(PD_{\mathrm{floor}}),$$

where

$$D_{\mathrm{floor}} := \widetilde{\Sigma}_\theta\Gamma(\widetilde{\Sigma}_\theta + \Gamma)^{-1} = U\,\mathrm{diag}(d_1, \ldots, d_p)U^\top,$$

with the d_i defined in the lemma statement. This follows by expanding the quadratic and completing the square in the positive-definite metric $\widetilde{\Sigma}_\theta + \Gamma$.

Let u_i be the ith column of U, and set $q_i := u_i^\top P u_i$. Then $0 \leq q_i \leq 1$ and $\sum_i q_i = \mathrm{rank}(P)$. Moreover,

$$\mathrm{tr}(PD_{\mathrm{floor}}) = \sum_{i=1}^{p} d_i q_i.$$

For attention, $\mathrm{rank}(P) = p - \mathrm{rank}(K)$, so Ky Fan's minimum principle gives the first inequality in (40); $\mathrm{rank}(K) \leq Hd$ gives the second. For the SSM, $\mathrm{rank}(P) \geq (p - Hd)_+$ gives (41). Summing over time and subtracting the Bayes baseline proves both finite-horizon bounds. At fixed $p, \widetilde{\Sigma}_\theta, \sigma^2$, all d_i are positive and $\mathcal{E}_T^{\mathrm{Bayes}} = o(T)$, so $Hd < p$ gives the claimed $\Omega(T)$ lower bounds. □

We now complete the proof of Proposition 1 by combining the kernel bounds with the architecture constructions. Fix $p, \Sigma_x, \Sigma_\theta, \sigma^2$ independently of T. For attention, Lemma 6 reduces the class lower bound to $\mathcal{E}_{L,T}$ up to $O(L/T) = O(1)$, with $1 \leq L \leq T$. Direct summation gives $TN_2 - N_1^2 \geq TL(L^2 - 1)/12$. Using also $N_1 \leq TL$ and $N_2 \leq LN_1$ in (27) yields

$$\mathcal{E}_{L,T} \geq \sum_{i=1}^{p} \left[\frac{\lambda_i(L^2 - 1)}{12(L + v_i/\lambda_i)} + \frac{v_i T}{L + v_i/\lambda_i}\right] \gtrsim L + \frac{T}{L}.$$

Since the Bayes subtraction is $O(\log T)$, any context achieving $O(\sqrt{T})$ regret must therefore satisfy $L = \Theta(\sqrt{T})$. On this range, (30) expands $\mathcal{E}_{L,T}$, so (28) gives the leading lower bound $\Lambda L/3 + VT/L \geq \frac{2}{\sqrt{3}}\bar{\zeta}\sqrt{T}$, minimized at the window (17). Thus, for every $d_e, d, H \geq 1$,

$$\inf_{L \geq 1} \mathrm{Reg}_T(\mathcal{F}_{\mathrm{att}}(1, d, H, L)) \geq \frac{2}{\sqrt{3}}\bar{\zeta}\sqrt{T} + o(\sqrt{T}) \tag{42}$$

uniformly over the class parameters. Whenever $d_e \geq 2p+1$ and $Hd \geq p$, the distributed-head uniform-kernel construction gives equality:

$$\mathrm{Reg}_T(\mathcal{F}_{\mathrm{att}}(1,d,H,L^\star(T))) = \frac{2}{\sqrt{3}}\bar{\zeta}\sqrt{T} + o(\sqrt{T}). \tag{43}$$

Outside this sufficient-width regime, only (42) is asserted. The bound on $\mathcal{E}_{L,T}$ also gives linear regret at fixed L, regardless of width.

For the SSM, Lemma 7 and Proposition 5 give, for every $d_e, d, H \geq 1$,

$$\mathrm{Reg}_T(\mathcal{F}_{\mathrm{SSM}}(1,d,H,1)) \geq \zeta\sqrt{T} + o(\sqrt{T}). \tag{44}$$

When $Hd \geq p$ and $d_e \geq 2p+2$, the modewise exponential construction attains this bound.

The aggregate-width claims for both classes follow from Lemma 8. When $Hd < p$, all d_i are strictly positive, and (21) yields the corresponding $\Omega(T)$ lower bounds. This completes the proof of Proposition 1.

Effect of task statistics. The leading regret coefficients distinguish how the two architectures respond to the task's statistical structure. Whenever both architecture witnesses are legal, their ratio is

$$\frac{2}{\sqrt{3}}\frac{\bar{\zeta}}{\zeta} = \frac{2}{\sqrt{3}}\frac{\sqrt{(\sum_i \lambda_i)(\sum_i v_i)}}{\sum_i \sqrt{\lambda_i v_i}} \geq \frac{2}{\sqrt{3}}. \tag{45}$$

Cauchy–Schwarz gives the inequality, with equality exactly when all λ_i are equal, equivalently $\Sigma_\theta = q\Sigma_x^{-1}$ for some $q > 0$.

Fix the total signal variance $\Lambda = \sum_i \lambda_i$, and the noise variance σ^2. Then $\bar{\zeta} = \sqrt{\Lambda((p+1)\Lambda + p\sigma^2)}$ is unchanged by redistributing signal variance across eigendirections. In contrast, $\zeta = \sum_i \sqrt{\lambda_i(\lambda_i + \Lambda + \sigma^2)}$ decreases when variance is transferred from a smaller eigenvalue to a larger one: the function $x \mapsto \sqrt{x(x+\Lambda+\sigma^2)}$ is strictly concave for $x > 0$. Thus, spectral concentration increases the SSM's advantage.

Optimal context and transition weights. For the full uniform and exponential lag profiles, the effective averaging width $(\sum_{u\geq 0}\kappa_u)^2/\sum_{u\geq 0}\kappa_u^2$ is the number of equally weighted evidence atoms giving the same innovation variance after normalizing the kernel mass to one. It equals L and $(1+\alpha)/(1-\alpha)$, respectively, as used in the proof of Lemma 3. Equations (17) and (19) give the leading-order optimal widths:

$$L^\star(T) = \sqrt{\frac{3VT}{\Lambda}} + O(1), \qquad \frac{1+\alpha_i^\star}{1-\alpha_i^\star} = 2\sqrt{\frac{v_i T}{\lambda_i}}.$$

More evidence-innovation variance relative to signal variance calls for longer averaging. Attention uses one shared window, which is unchanged by spectral concentration at fixed total signal variance and noise; the SSM instead adjusts each decay to its eigendirection. In the p-dimensional diagonal realization of Lemma 7, $\|\Phi^\star\|_{\mathrm{op}} = \max_i \alpha_i^\star = 1 - T^{-1/2}\min_i\sqrt{\lambda_i/v_i} + O(T^{-1})$. Thus, as the horizon increases, the SSM realizes longer memory through its spectral radius approaching one at rate $T^{-1/2}$, with a leading constant determined by the signal-to-innovation variance ratios.

Loss landscape at the optimal weights. The optimal weights also reveal how the local optimization landscape changes with the horizon. We compare average-loss curvatures along selected weight directions because larger curvature makes gradient updates more sensitive to the step size. In the diagonal SSM construction of Lemma 7, fix an eigendirection i and vary its transition weight α and read gain β, keeping the other maps fixed. Its kernel is $\kappa_{u,i} = \beta\alpha^u$, and Lemma 1 gives its cumulative contribution to belief-estimation error:

$$J_{T,i}^{\mathrm{SSM}}(\alpha,\beta) = \sum_{t=1}^{T}\left\{\lambda_i\left[\frac{\beta(1-\alpha^t)}{1-\alpha} - 1\right]^2 + \frac{v_i\beta^2(1-\alpha^{2t})}{1-\alpha^2}\right\}.$$

At the leading-order optimal construction, $\alpha = \alpha_i^\star$ and $\beta = 1-\alpha_i^\star$. Differentiating with the read gain held fixed gives

$$\frac{1}{T}\frac{\partial^2 J_{T,i}^{\mathrm{SSM}}}{\partial\alpha^2}\bigg|_{(\alpha,\beta)=(\alpha_i^\star,\,1-\alpha_i^\star)} = \frac{2\lambda_i}{(1-\alpha_i^\star)^2} + O(\sqrt{T}) = 2v_iT + O(\sqrt{T}).$$

This is also the transition-coordinate curvature of the average prediction loss and regret, since the omitted terms are independent of these weights. Thus, although the optimal transition weights remain bounded, the transition-weight curvature grows linearly with the horizon, with a leading coefficient proportional to the evidence-innovation variance v_i.

For comparison, in the uniform-attention construction of Lemma 5, vary the query-projection gain a in eigendirection i, keeping the other maps fixed. Recalling $\tau_t = \min(t, L)$ and $V = \sum_j v_j$, the quadratic in the proof of Lemma 6 gives

$$J^{\mathrm{att}}_{T,i}(a; L) = \sum_{t=1}^{T} \left[\lambda_i (a\tau_t - 1)^2 + v_i a^2 \tau_t\right].$$

At $L = L^\star(T)$, its average-loss curvature is

$$\frac{1}{T}\frac{\partial^2 J^{\mathrm{att}}_{T,i}}{\partial a^2}\bigg|_{L=L^\star(T)} = \frac{2}{T}\sum_{t=1}^{T}(\lambda_i \tau_t^2 + v_i \tau_t) = \frac{6\lambda_i V}{\Lambda} T + O(\sqrt{T}).$$

Both curvatures grow linearly with the horizon. In the specified query-gain and transition-weight coordinates, the leading coefficients sum to $6V$ for attention and $2V$ for the SSM. Thus, local step-size sensitivity increases with the horizon. Attention's summed leading curvature is three times the SSM's in these coordinates, but this does not establish a convergence-rate ordering between the families.

E.2 Extension to time- and data-adaptive kernels

This section shows how count normalization and additional depth let both families adapt their kernels to elapsed time and realized information, escaping the stationary regret floor of Theorem 1. We first explain the two mechanisms in the scalar case (§E.2.1), then extend the predictors and their regret rates to general p and state the architecture-class optima (§E.2.2). Sections E.2.3 and E.2.4 prove the matching lower bounds and give the time- and data-adaptive constructions, respectively.

E.2.1 Scalar adaptive kernels and regret rates

To make the two adaptive mechanisms clear, we first consider the scalar case $p = 1$, with $\Sigma_x = 1$ and $\Sigma_\theta = \nu > 0$. Then $A_t = \sum_{s\le t} x_s y_s$, $B_t = \sum_{s\le t} x_s^2$, and $\hat\theta_t^{\mathrm{Bayes}} = \nu A_t/(\sigma^2 + \nu B_t)$.

As in Lemma 1, kernel mismatch has a bias–variance decomposition. Here we condition on the observed regressors to allow data-dependent kernel weights. For a predictor g, write its projected belief as $\Pi_t\hat\theta_t(g) = \sum_{s\le t} k_{t,s}(g) x_s y_s$. As in the definition of $\mathcal{K}_t$, these kernel weights depend only on the regressors $x_{1:t}$. Define $\mu_t(g) := \sum_{s\le t} k_{t,s}(g) x_s^2$ and $V_t(g) := \sum_{s\le t} k_{t,s}(g)^2 x_s^2$. Here $\mu_t(g)$ is the regressor-weighted kernel mass, whose deviation from one determines the bias, and $\sigma^2 V_t(g)$ is the conditional observation-noise variance.

Lemma 9 (Scalar regret decomposition)**.** *For any class $\mathcal{F}$, $\mathrm{Reg}_T(\mathcal{F}) = \inf_{g\in\mathcal{F}}\{\mathsf{K}_T(g) + \mathsf{R}_T(g)\}$, where*

$$\mathsf{R}_T(g) := \sum_{t\le T} \left\| g_t - x_{t+1}\Pi_t\hat\theta_t(g)\right\|_t^2$$

charges departures from kernel-form belief prediction, and

$$\begin{aligned}\mathsf{K}_T(g) &:= \sum_{t\le T}\mathbb{E}\Big[\big(\Pi_t\hat\theta_t(g) - \hat\theta_t^{\mathrm{Bayes}}\big)^2\Big] \\ &= \nu\sum_{t\le T}\mathbb{E}[(\mu_t(g)-1)^2] + \sigma^2\sum_{t\le T}\mathbb{E}[V_t(g)] - \sigma^2\log T + O(1).\end{aligned}$$

Proof. The Pythagorean identity (2) separates kernel mismatch from the two orthogonal residual terms, giving the first identity. To evaluate the kernel mismatch, condition on $x_{1:t}$ and substitute $y_s = x_s\theta + \epsilon_s$. The projected belief is $\mu_t(g)\theta + \sum_{s\le t} k_{t,s}(g) x_s\epsilon_s$, so independence of θ and the observation noise gives

$$\mathbb{E}\Big[(\Pi_t\hat\theta_t(g) - \theta)^2 \mid x_{1:t}\Big] = \nu(\mu_t(g)-1)^2 + \sigma^2 V_t(g).$$

The identity extends from bounded kernel weights to the closure defining $\mathcal{K}_t$ by L^2 approximation. Since the Bayes belief is the conditional mean of θ, posterior orthogonality then yields

$$\mathbb{E}[(\Pi_t\hat{\theta}_t(g) - \hat{\theta}_t^{\mathrm{Bayes}})^2] = \nu\mathbb{E}[(\mu_t(g)-1)^2] + \sigma^2\mathbb{E}[V_t(g)] - \mathbb{E}\left[\frac{\nu\sigma^2}{\sigma^2+\nu B_t}\right].$$

Summing and using the scalar specialization of (21) proves the claim. □

The final term subtracts the cumulative uncertainty already incurred by the Bayes predictor. To motivate adaptive kernels, write $U_t := B_t/t$. The Bayes kernel satisfies

$$k_{t,s}^{\mathrm{Bayes}} = \frac{1}{t\big(U_t + \sigma^2/(\nu t)\big)} \sim \frac{1}{tU_t} \qquad (t\to\infty).$$

The factor $1/t$ adapts to elapsed time, while $1/U_t$ adjusts for realized information. Since U_t concentrates at 1, this suggests kernels

$$k_{t,s} = \frac{\hat{\mu}(U_t)}{t},$$

where $\hat{\mu}(v)$ approximates $1/v$ near $v=1$. The next theorem quantifies how the accuracy of this local approximation determines cumulative kernel mismatch.

Theorem 4 (Adaptive rate)**.** *Suppose g realizes $k_{t,s} = \hat{\mu}(U_t)/t$ for a fixed measurable function $\hat{\mu}$ of polynomial growth on $[0,\infty)$, meaning $|\hat{\mu}(v)| \le C(1+v^m)$ for some $C>0$ and $m\ge 0$. Assume $1 - v\hat{\mu}(v) = c(v-1)^r + O(|v-1|^{r+\eta})$ as $v\to 1$, for some integer $r\ge 1$, $c\neq 0$, and $\eta>0$. Then*

$$\mathsf{K}_T(g) = \begin{cases} 2\nu c^2 \log T + O(1), & r=1,\\ \Theta(1), & r\ge 2. \end{cases}$$

Proof. The variance has the same leading term as Bayes, so the remaining bias determines the rate. For the displayed kernel,

$$\mu_t(g) = U_t\hat{\mu}(U_t), \qquad V_t(g) = \frac{U_t\hat{\mu}(U_t)^2}{t}.$$

Since $B_t \sim \chi^2_t$, $\mathbb{E}[(U_t-1)^2] = 2/t$ and $\mathbb{E}|U_t-1|^q = O(t^{-q/2})$ for every fixed $q>0$. For even moments, the latter bound follows by expanding powers of $\sum_{s\le t}(x_s^2-1)$: every nonzero term repeats each index at least twice. Hölder's inequality gives the remaining moments. The exponential tails of U_t, together with polynomial growth and Cauchy–Schwarz, make contributions outside any fixed neighborhood of 1 exponentially small. We may therefore average the local matching expansion to obtain

$$\mathbb{E}[(\mu_t(g)-1)^2] = \begin{cases} 2c^2/t + O(t^{-1-\eta/2}), & r=1,\\ O(t^{-r}), & r\ge 2. \end{cases}$$

Likewise, $v\hat{\mu}(v)^2 = 1 + O(|v-1|)$ near 1, hence $\mathbb{E}[U_t\hat{\mu}(U_t)^2] = 1 + O(t^{-1/2})$. Thus the variance term cancels the Bayes baseline up to a bounded remainder:

$$\sigma^2\sum_{t\le T}\mathbb{E}[V_t(g)] - \sigma^2\log T = \sigma^2\sum_{t\le T}\frac{1+O(t^{-1/2})}{t} - \sigma^2\log T = O(1).$$

Lemma 9 now gives the logarithmic formula for $r=1$ and the bounded upper bound for $r\ge 2$.

For the positive lower bound, it suffices to consider $t=1$. The first-step Bayes kernel is $1/(v+\sigma^2/\nu)$, whereas the matching assumption implies $\hat{\mu}(v)\to 1$ as $v\to 1$. These kernels differ on an interval around 1, which $U_1 = x_1^2$ visits with positive probability. The first-step kernel mismatch is therefore strictly positive; all later terms are nonnegative. This proves $\Theta(1)$ for $r\ge 2$. □

The constant approximation $\hat{\mu}(v)=1$ and the first-order Taylor approximation $\hat{\mu}(v) = 2-v$ give the predictors

$$g_t^{\mathrm{time}} := x_{t+1}\frac{A_t}{t}, \qquad g_t^{\mathrm{data}} := x_{t+1}(2-U_t)\frac{A_t}{t}.$$

Both apply kernel-form beliefs to x_{t+1}, so their residual costs vanish and these kernel costs are their full regrets. For g^{time}, $1 - v\hat{\mu}(v) = -(v-1)$ gives $\mathrm{Reg}_T(g^{\mathrm{time}}) = 2\nu\log T + O(1)$. The data-adaptive correction cancels the first-order fluctuation, since

$$(2-U_t)U_t = 1 - (U_t-1)^2.$$

Thus g^{data} has matching order two and $\mathrm{Reg}_T(g^{\mathrm{data}}) = \Theta(1)$.

E.2.2 Adaptive kernels and class optima for general p

The same correction extends to any fixed p. In the whitened coordinates of Appendix E.1, define the normalized evidence vector a_t and information matrix U_t by

$$a_t := \frac{\Sigma_x^{-1/2} A_t}{t} = \frac{1}{t}\sum_{s\le t} \widetilde{x}_s y_s, \qquad U_t := \frac{\widetilde{B}_t}{t} = \frac{1}{t}\sum_{s\le t} \widetilde{x}_s \widetilde{x}_s^\top.$$

Count normalization again uses a_t as the belief. The scalar cancellation extends directly to

$$(2I_p - U_t)U_t = I_p - (U_t - I_p)^2,$$

so the same first-order cancellation applies to the coefficient on $\widetilde{\theta}$. The corresponding predictors are

$$g_t^{\mathrm{time}} := \widetilde{x}_{t+1}^\top a_t, \qquad g_t^{\mathrm{data}} := \widetilde{x}_{t+1}^\top (2I_p - U_t) a_t.$$

Lemma 10 gives the corresponding regret bounds for general p.

Lemma 10 (Regret of adaptive predictors)**.** *For every $t \ge 1$,*

$$\mathbb{E}\|a_t - \widetilde{\theta}\|_2^2 = \frac{\operatorname{tr}\Gamma}{t}, \qquad \mathbb{E}\|(2I_p - U_t)a_t - \widetilde{\theta}\|_2^2 = \frac{p\sigma^2}{t} + O(t^{-2}).$$

The prediction losses are these quantities plus σ^2. Both predictors have zero residual cost in (2), and

$$\mathrm{Reg}_T(g^{\mathrm{time}}) = (p+1)\operatorname{tr}(\widetilde{\Sigma}_\theta)\log T + O(1), \qquad \mathrm{Reg}_T(g^{\mathrm{data}}) = O(1).$$

Proof. Count normalization leaves the empirical information fluctuation intact; the correction cancels it to first order. For the first error, the innovation covariance calculation in Lemma 1 gives

$$\mathbb{E}[(\widetilde{x}_r y_r - \widetilde{\theta})(\widetilde{x}_s y_s - \widetilde{\theta})^\top] = \mathbf{1}\{r = s\}\Gamma.$$

Averaging these innovations gives $\operatorname{tr}(\Gamma)/t$.

For the correction, put $\Delta_t := U_t - I_p$. The exact identity $(2I_p - U_t)U_t = I_p - \Delta_t^2$ gives

$$(2I_p - U_t)a_t - \widetilde{\theta} = -\Delta_t^2\widetilde{\theta} + (I_p - \Delta_t)\frac{1}{t}\sum_{s\le t}\widetilde{x}_s\epsilon_s.$$

Conditioned on $\widetilde{x}_{1:t}$, the two terms are independent and centered, and $t^{-1}\sum_{s\le t}\widetilde{x}_s\epsilon_s$ has covariance $\sigma^2 U_t/t$. Taking the squared norm and averaging therefore gives

$$\mathbb{E}\|(2I_p - U_t)a_t - \widetilde{\theta}\|_2^2 = \mathbb{E}\operatorname{tr}(\widetilde{\Sigma}_\theta\Delta_t^4) + \frac{\sigma^2}{t}\mathbb{E}\operatorname{tr}(I_p - \Delta_t - \Delta_t^2 + \Delta_t^3).$$

Each entry of Δ_t is an average of independent centered variables $\widetilde{x}_{s,i}\widetilde{x}_{s,j} - \mathbf{1}\{i=j\}$. In its fourth-moment expansion, each nonzero term repeats every sample index at least twice. There are $O(t^2)$ such terms with bounded moments, so $\mathbb{E}|\Delta_{t,ij}|^4 = O(t^{-2})$. Since p is fixed, summing entries and applying Hölder's inequality gives

$$\mathbb{E}\|\Delta_t\|_F^k = O(t^{-k/2}), \qquad k = 2,3,4.$$

Together with $\mathbb{E}\Delta_t = 0$, these bounds make the bias term $O(t^{-2})$ and the noise term $p\sigma^2/t + O(t^{-2})$. They hold from $t = 1$, including $t < p$, without an empirical inverse.

Both predictions are linear in the fresh regressor. Their original-coordinate kernels are Σ_x^{-1}/t and $\Sigma_x^{-1/2}(2I_p - U_t)\Sigma_x^{-1/2}/t$, respectively. Their weights depend only on the regressors; truncating them and using finite Gaussian moments shows that both beliefs belong to the closure defining $\mathcal{K}_t$. Both residual terms in (2) therefore vanish. Finally, (7) subtracts the Bayes baseline (21). Since $\operatorname{tr}\Gamma - p\sigma^2 = (p+1)\operatorname{tr}\widetilde{\Sigma}_\theta$ and the $O(t^{-2})$ remainders are summable, this proves both regret bounds. □

The following propositions establish the optima of the count-normalized classes (Appendix C.2). Sections E.2.3 and E.2.4 prove the matching class lower bounds and construct both architectures to realize g^{time} and g^{data}, respectively.

Proposition 6 (Time-adaptive pair). *For $d, H \geq 1$ with $Hd \geq p$ and $d_e \geq 2p+2$, each of*

$$\mathcal{F}^{\mathrm{cal}}_{\mathrm{att}}(1, d, H, T; t), \qquad \mathcal{F}^{\mathrm{cal}}_{\mathrm{SSM}}(1, d, H, 1; t)$$

contains g^{time}. For either class $\mathcal{F}$,

$$\mathrm{Reg}_T(\mathcal{F}) = (p+1)\,\mathrm{tr}(\widetilde{\Sigma}_\theta)\log T + O(1).$$

One additional layer makes the kernel data-adaptive through the correction $2I_p - U_t$. Attention implements this correction on the fresh regressor, whereas the SSM implements it on the maintained belief.

Proposition 7 (Data-adaptive pair). *The predictor g^{data} is realized under the following sufficient conditions:*

(a) Attention $\mathcal{F}^{\mathrm{cal}}_{\mathrm{att}}(2, d, H, T; (t, t))$ with $d \geq p$, $H \geq 1$, and $d_e \geq 3p+2$.

(b) The fixed-transition, ordinary-read class $\mathcal{F}^{\mathrm{cal}}_{\mathrm{SSM}}(2, d, H, 1; (t, 1))$ with $d, H \geq 1$, $Hd \geq p^2 + p$, and $d_e \geq p^2 + 3p + 2$.

Under these conditions, either class satisfies $\mathrm{Reg}_T(\mathcal{F}) = \Theta(1)$.

Thus time adaptation through count normalization gives both families logarithmic regret, while data adaptation to realized information reduces it to constant regret.

E.2.3 Proof of Proposition 6: Time-adaptive pair

Count normalization changes the coefficients of a one-layer predictor, but not its polynomial degree. Lemma 11 establishes this restriction, and Lemma 12 turns it into a logarithmic regret floor. We then construct both architectures to attain it. We use the observed polynomial space $\mathcal{Q}_{2,t}$ and deterministic kernel-form space $\mathcal{S}_t$ from §E.1.3.

Lemma 11 (Degree-two implicit beliefs). *For every $d_e, d, H \geq 1$, the whitened belief $\widetilde{\theta}_t(g)$ belongs to $\mathcal{Q}_{2,t}$ whenever*

$$g \in \mathcal{F}^{\mathrm{cal}}_{\mathrm{att}}(1, d, H, T; t) \quad \text{or} \quad g \in \mathcal{F}^{\mathrm{cal}}_{\mathrm{SSM}}(1, d, H, 1; t).$$

The conclusion is uniform in the architecture parameters, even when they depend on T.

Proof. Each attention summand is a bilinear score times an affine value, hence has degree at most three in the tokens. For a fixed-transition SSM, every state coordinate is a linear combination of bilinear writes, and the ordinary read likewise gives degree at most three. Affine biases, positional biases, residual maps, and deterministic count normalization do not increase this bound.

In $\widetilde{\theta}_t(g) = \mathbb{E}[\widetilde{x}_{t+1} g_t \mid \mathcal{G}_t^-]$, a surviving monomial must contain a fresh-regressor factor before multiplication by $\widetilde{x}_{t+1}$. Gaussian integration removes these factors and leaves observed degree at most two. Overlap between neighboring token blocks does not increase the degree, and widths and parameter magnitudes affect only the coefficients. □

Lemma 12 (Degree-two regret floor). *If $\widetilde{\theta}_t(g) \in \mathcal{Q}_{2,t}$ at every step, then*

$$\mathrm{Reg}_T(g) \geq \sum_{t \leq T} \sum_{i=1}^{p} \frac{\lambda_i v_i}{t\lambda_i + v_i} - \mathcal{E}_T^{\mathrm{Bayes}} = (p+1)\,\mathrm{tr}(\widetilde{\Sigma}_\theta)\log T + O(1).$$

The bound is uniform in g, including horizon-dependent parameters.

Proof. Lemma 4 reduces the best degree-two approximation of $\widetilde{\theta}$ to its projection onto $\mathcal{S}_t$. The evidence moments from Lemma 1 are

$$\mathbb{E}[\widetilde{\theta}(\widetilde{x}_s y_s)^\top] = \widetilde{\Sigma}_\theta, \qquad \mathbb{E}[(\widetilde{x}_r y_r)(\widetilde{x}_s y_s)^\top] = \widetilde{\Sigma}_\theta + \mathbf{1}\{r = s\}\Gamma.$$

Thus the projection is

$$\widetilde{\theta}_t^{(2)} = \widetilde{\Sigma}_\theta (t\widetilde{\Sigma}_\theta + \Gamma)^{-1} \sum_{s \leq t} \widetilde{x}_s y_s,$$

since its error is orthogonal to every evidence atom. Using the eigenvalues λ_i, v_i from (4), its mean-squared error is

$$\mathbb{E}\|\widetilde{\theta} - \widetilde{\theta}_t^{(2)}\|_2^2 = \sum_{i=1}^{p} \frac{\lambda_i v_i}{t\lambda_i + v_i} = \frac{\operatorname{tr}\Gamma}{t} + O(t^{-2}).$$

Equation (22) charges other polynomial components nonnegatively, as does (7) for the fresh-regressor residual. Summing and subtracting the Bayes baseline (21) proves the claim, since $\operatorname{tr}\Gamma - p\sigma^2 = (p+1)\operatorname{tr}\widetilde{\Sigma}_\theta$. □

Combining Lemmas 11 and 12 gives the lower bounds in Proposition 6. To attain them, preserve $(\widetilde{x}_{t+1}, \widetilde{x}_t, y_t)$ in $2p+1$ residual coordinates and reserve one clean prediction coordinate. For attention, partition $[p]$ among at most H sets I_b of size at most d. Head b uses query $\widetilde{x}_{t+1,I_b}$, key $\widetilde{x}_{s,I_b}$, and value y_s in one value coordinate, with unused slots and positional biases zeroed. Output maps sum the count-normalized heads into the prediction coordinate, giving

$$g_t = \frac{1}{t}\sum_{s\le t}\sum_{b} \widetilde{x}_{t+1,I_b}^{\top}\widetilde{x}_{s,I_b}y_s = \widetilde{x}_{t+1}^{\top}a_t = g_t^{\mathrm{time}}.$$

For the SSM, distribute p state coordinates across the heads. Coordinate i has transition 1, write $\widetilde{x}_{t,i}y_t$, and read selector $\widetilde{x}_{t+1,i}$. Its state is the unnormalized sum $\sum_{s\le t}\widetilde{x}_{s,i}y_s$; normalization divides only its read by t. Summing these reads into the clean prediction coordinate gives the same output. Both constructions fit whenever $Hd \ge p$ and $d_e \ge 2p+2$, and Lemma 10 gives their regret with zero residual cost. The SSM uses p active states out of Hd.

E.2.4 Proof of Proposition 7: Data-adaptive pair

A positive first-step excess loss supplies the lower bound for constant regret. Lemma 13 establishes this floor; we then construct both two-layer predictors to attain constant regret. At the first prediction time, the token is $z_1 = (x_2, y_1, x_1)$ and

$$g_1^{\mathrm{Bayes}}(z_1) = \frac{y_1 x_2^{\top}\Sigma_\theta x_1}{\sigma^2 + x_1^{\top}\Sigma_\theta x_1}.$$

Lemma 13 (Positive first-step regret)**.** *Every predictor g in either class*

$$\mathcal{F}_{\mathrm{att}}^{\mathrm{cal}}(2,d,H,T;(t,t)), \qquad \mathcal{F}_{\mathrm{SSM}}^{\mathrm{cal}}(2,d,H,d_{\mathrm{conv}};(t,1))$$

satisfies

$$\mathrm{Reg}_T(g) \ge c > 0,$$

where c depends only on the fixed distributional parameters. The bound is uniform over $T, d_e, d, H, d_{\mathrm{conv}} \ge 1$.

Proof. At $t = 1$, both normalization schedules equal $(1,1)$. An attention layer's bilinear score times its affine value increases polynomial degree by at most a factor of three. For the SSM, zero padding and zero initial state leave only a bilinear current-token write followed by a token–state read, giving the same degree bound regardless of convolution width. Affine maps and residuals do not increase this bound. Thus every two-layer first prediction belongs to the space $\mathcal{P}_9$ of polynomials of total degree at most nine in z_1, independently of widths or parameter values.

The law of z_1 has an everywhere-positive density and finite polynomial moments. If a polynomial P equaled the Bayes predictor almost surely, multiplying by its denominator would give a polynomial identity. Fix $v \ne 0$, set $x_1 = rv$, $x_2 = v$, $y_1 = 1$, and write $\alpha = v^{\top}\Sigma_\theta v > 0$. The identity would imply

$$(\sigma^2 + \alpha r^2)P(v, 1, rv) = \alpha r,$$

which is impossible for a univariate polynomial. Since $\mathcal{P}_9$ is finite-dimensional and therefore closed in $L^2(P_{z_1})$, its squared distance from the Bayes predictor is positive:

$$c := \inf_{P\in\mathcal{P}_9} \mathbb{E}[(P(z_1) - g_1^{\mathrm{Bayes}}(z_1))^2] > 0.$$

Bayes orthogonality and nonnegative later excess losses give $\mathrm{Reg}_T(g) \ge \mathbb{E}[(g_1 - g_1^{\mathrm{Bayes}})^2] \ge c$, including for horizon-dependent widths and parameters. □

It remains to realize g^{data} in each architecture class. For attention, preserve the $2p+1$ raw whitened token coordinates and reserve a zero-initialized p-vector block and scalar output coordinate. One active head per layer suffices when $d \geq p$. The first layer uses query $\widetilde{x}_{t+1}$, key $\widetilde{x}_s$, and value $\widetilde{x}_s$, writing only into the vector block:

$$\frac{1}{t}\sum_{s\leq t}(\widetilde{x}_{t+1}^\top \widetilde{x}_s)\widetilde{x}_s = U_t \widetilde{x}_{t+1}.$$

The second layer uses the affine query $2\widetilde{x}_{t+1} - U_t\widetilde{x}_{t+1}$, selected from the raw and new blocks, and retains key $\widetilde{x}_s$ and value y_s from the untouched raw source coordinates. Its output in the clean scalar coordinate is

$$\frac{1}{t}\sum_{s\leq t}(2\widetilde{x}_{t+1} - U_t\widetilde{x}_{t+1})^\top \widetilde{x}_s y_s = \widetilde{x}_{t+1}^\top(2I_p - U_t)a_t = g_t^{\mathrm{data}}.$$

Here U_t is symmetric. Both layers use full context, normalization by t, and zero positional biases. Padding and zeroing unused heads extend the construction to every $d \geq p$ and $d_e \geq 3p+2$.

The SSM construction uses two fixed-transition layers with ordinary reads and identity convolutions. Alongside the $2p+1$ raw token coordinates, reserve residual blocks for a_t, all entries of U_t, and the prediction. Its first layer uses p recurrent coordinates to accumulate evidence and p^2 to accumulate information, with identity transitions and writes $\widetilde{x}_{t,j}y_t$ and $\widetilde{x}_{t,i}\widetilde{x}_{t,j}$, respectively. They hold the unnormalized statistics

$$(\Sigma_x^{-1/2}A_t)_j, \qquad (\Sigma_x^{-1/2}B_t\Sigma_x^{-1/2})_{ij}, \qquad i,j \in [p].$$

Constant read selectors 1 with normalization by t expose $a_{t,j}$ and $(U_t)_{ij}$ in their reserved residual blocks; normalization does not alter the stored recurrences.

The second layer has one state coordinate per ordered pair (i,j), each with transition zero and normalization factor 1. Its affine write factors are $a_{t,j}$ and $2\mathbf{1}\{i=j\} - (U_t)_{ij}$; the ordinary read selector is $\widetilde{x}_{t+1,i}$. Summing these coordinatewise reads into the clean output gives

$$\sum_{i,j}\widetilde{x}_{t+1,i}a_{t,j}\big(2\mathbf{1}\{i=j\} - (U_t)_{ij}\big) = \widetilde{x}_{t+1}^\top(2I_p - U_t)a_t = g_t^{\mathrm{data}}.$$

All constants are legal affine biases, and the upper layer accesses the statistics only through residual coordinates. Every state has its own affine write and read rows, so packing at most d coordinates per head is valid whenever $Hd \geq p^2+p$. The residual width required is $d_e \geq (2p+1)+p+p^2+1 = p^2+3p+2$. The construction uses p^2+p active states in the first layer and p^2 in the second, totaling $2p^2+p$ out of $2Hd$ across both layers.

Both constructions realize g^{data} at every prediction time, including startup. Lemma 10 gives $O(1)$ regret with zero residual cost, and Lemma 13 supplies the positive class floor. This proves Proposition 7.

F Case 2: full statements and proofs

For an overview of the notation and proof dependencies, see the notation guide and proof guide at the start of the appendix.

We first prove the two-factor floor (§F.1). Section F.2 develops Case 2.a: assembly floors, constructions, class lower bounds, and the memory comparison. Section F.3 covers Case 2.b: addressing-capacity bounds and both constructions.

F.1 Proof of Theorem 2

We use the whitened coordinates of §E.1.1. We extend the same belief projection to routing and use it to prove the two-factor floor in Theorem 2. Recall the whitened coordinates

$$\widetilde{x}_t := \Sigma_x^{-1/2}x_t, \qquad \widetilde{\theta} := \Sigma_x^{1/2}\theta, \qquad \widetilde{\Sigma}_\theta := \Sigma_x^{1/2}\Sigma_\theta\Sigma_x^{1/2},$$

so that $\widetilde{x}_t \sim \mathcal{N}(0, I_p)$ and $y_s = \widetilde{x}_{\rho(s)}^\top\widetilde{\theta} + \epsilon_s$ at a matched step. Throughout this appendix, the route is either deterministic, as in positional routing, or is measurable with respect to an auxiliary key process

independent of $(\widetilde{\theta}, (\widetilde{x}_t)_t, (\epsilon_t)_t)$, as in content routing. Together with partial injectivity, this makes the routed regressors at distinct matched steps distinct iid standard Gaussian vectors.

Recall the Case 1 notation

$$\widetilde{\Sigma}_\theta = U \operatorname{diag}(\lambda_1, \dots, \lambda_p) U^\top, \qquad v_i := \lambda_i + \operatorname{tr} \widetilde{\Sigma}_\theta + \sigma^2,$$

and

$$\Lambda := \operatorname{tr} \widetilde{\Sigma}_\theta, \qquad V := \sum_{i=1}^{p} v_i, \qquad d_i := \frac{\lambda_i v_i}{\lambda_i + v_i}.$$

For n distinct matched observations, let

$$P_n := \left(\widetilde{\Sigma}_\theta^{-1} + \sigma^{-2} \sum_{j=1}^{n} \widetilde{x}_j \widetilde{x}_j^\top \right)^{-1}, \qquad \mathcal{E}_N^{\mathrm{Bayes}} = \sum_{n=1}^{N} \mathbb{E} \operatorname{tr} P_n. \tag{46}$$

At fixed p and fixed $\widetilde{\Sigma}_\theta \succ 0$,

$$\mathcal{E}_N^{\mathrm{Bayes}} = p\sigma^2 \log N + O(1). \tag{47}$$

Here $\mathbb{E} \operatorname{tr} P_n = e_n^{\mathrm{Bayes}}$ from (8), so (47) is the Case 1 Bayes error estimate after N matched observations.

For each time t, write

$$\mathcal{M}_t := \{s \le t : s \text{ is matched}\}, \qquad N_t^\rho := |\mathcal{M}_t|.$$

Fix a prediction time t for which $t+1$ is matched. Let $\mathcal{G}_t^-$ contain the route, all observed labels, and every regressor in the history except all token occurrences of $\widetilde{x}_{\rho(t+1)}$. Although this regressor can appear in earlier tokens, exogenous and injective routing ensures that no observed label uses it. Thus

$$\widetilde{x}_{\rho(t+1)} \sim \mathcal{N}(0, I_p) \quad \text{and} \quad \widetilde{x}_{\rho(t+1)} \text{ is independent of } \mathcal{G}_t^-.$$

For an arbitrary square-integrable prediction g_t, the Case 1 projection (5) therefore becomes

$$\widetilde{\theta}_t(g) := \mathbb{E}[\widetilde{x}_{\rho(t+1)} g_t \mid \mathcal{G}_t^-], \qquad g_t^\perp := g_t - \widetilde{x}_{\rho(t+1)}^\top \widetilde{\theta}_t(g). \tag{48}$$

The Case 1 projection argument in §E.1.1 shows that the map $a \mapsto \widetilde{x}_{\rho(t+1)}^\top a$ is an isometry from $L^2(\mathcal{G}_t^-; \mathbb{R}^p)$ into scalar L^2. Equation (48) is the orthogonal projection of g_t onto its image. The Bayes belief $\widetilde{\theta}_t^{\mathrm{Bayes}} := \mathbb{E}[\widetilde{\theta} \mid z_{1:t}]$ is $\mathcal{G}_t^-$-measurable because no observed label uses the held-out regressor. Thus

$$g_t^{\mathrm{Bayes}} = \widetilde{x}_{\rho(t+1)}^\top \widetilde{\theta}_t^{\mathrm{Bayes}}, \tag{49}$$

$$\|g_t - g_t^{\mathrm{Bayes}}\|_t^2 = \|\widetilde{\theta}_t(g) - \widetilde{\theta}_t^{\mathrm{Bayes}}\|_t^2 + \|g_t^\perp\|_t^2, \tag{50}$$

where, for vector-valued beliefs, $\langle a, b \rangle_t := \mathbb{E}[a^\top b]$ and $\|a\|_t^2 := \mathbb{E}\|a\|_2^2$.

For vectors in either Hilbert space, write

$$\operatorname{corr}^2(A, B) := \frac{\langle A, B \rangle^2}{\|A\|^2 \|B\|^2}.$$

Throughout the Case 2 proofs, we use the convention $0/0 = 0$.

Proof of Theorem 2. At a matched target, distance to the particular prediction g_t dominates distance to the line it spans:

$$\|g_t - g_t^{\mathrm{Bayes}}\|_t^2 \ge \inf_{c \in \mathbb{R}} \|c g_t - g_t^{\mathrm{Bayes}}\|_t^2 = \|g_t^{\mathrm{Bayes}}\|_t^2 \big(1 - \operatorname{corr}^2(g_t, g_t^{\mathrm{Bayes}})\big).$$

By orthogonality in (48) and the Bayes factorization (49),

$$\langle g_t, g_t^{\mathrm{Bayes}} \rangle_t = \langle \widetilde{\theta}_t(g), \widetilde{\theta}_t^{\mathrm{Bayes}} \rangle_t, \qquad \|g_t^{\mathrm{Bayes}}\|_t^2 = \|\widetilde{\theta}_t^{\mathrm{Bayes}}\|_t^2,$$
$$\|g_t\|_t^2 = \|\widetilde{\theta}_t(g)\|_t^2 + \|g_t^\perp\|_t^2.$$

Consequently,

$$\operatorname{corr}^2(g_t, g_t^{\mathrm{Bayes}}) = \operatorname{corr}^2(\widetilde{\theta}_t(g), \widetilde{\theta}_t^{\mathrm{Bayes}}) \frac{\|\widetilde{\theta}_t(g)\|_t^2}{\|\widetilde{\theta}_t(g)\|_t^2 + \|g_t^\perp\|_t^2}.$$

Unmatched targets contribute a nonnegative excess loss, so summing over matched targets proves the theorem. □

F.2 Case 2.a: positional routing

We first prove the assembly floors of Corollary 1 (§F.2.1), then give the constructions and upper bounds for Proposition 2 (§F.2.2). The attention and SSM lower bounds for Proposition 2 follow in §F.2.3 and §F.2.4, respectively. The attention bound also proves Proposition 3, and we conclude with the memory comparison.

We now specialize the geometry of Appendix F.1 to the deterministic route $\rho(s) = s - K$, with $T_\rho = T - K$. We retain the whitened coordinates and innovation covariance Γ from §E.1.1. To reuse the Case 1 calculations, we index matched observations by $j = s - K$ and write $n = t - K$ for the number available at prediction time t:

$$Y_j := y_{K+j} = \widetilde{x}_j^\top \widetilde{\theta} + \epsilon_{K+j}, \qquad r_j := \widetilde{x}_j Y_j, \qquad 1 \le j \le T_\rho.$$

After n matched observations, the Bayes belief is

$$\widetilde{\theta}^{\mathrm{Bayes}}_{K+n} = \mathbb{E}[\widetilde{\theta} \mid \widetilde{x}_{1:n}, Y_{1:n}]. \tag{51}$$

At prediction time $t = K + n$, the routed regressor for the next target is $\widetilde{x}_{n+1}$ and the evidence atoms $r_1, \dots, r_n$ are available. The single matched prediction before the first observation has zero Bayes mean and contributes zero to each construction below, so the T_ρ rows $n = 1, \dots, T_\rho$ cover the nontrivial matched predictions.

The next lemma extends the Case 1 bias–variance decomposition to routed evidence, using the kernel mass M_n and innovation covariance Γ.

Lemma 14 (Exact estimation error and regret of a routed kernel)**.** *For deterministic matrices $K_{n,j} \in \mathbb{R}^{p\times p}$, the corresponding kernel-form predictor g has*

$$\widetilde{\theta}_{K+n}(g) = \sum_{j=1}^{n} K_{n,j} r_j, \qquad M_n := \sum_{j=1}^{n} K_{n,j}.$$

Then

$$\mathbb{E}\|\widetilde{\theta}_{K+n}(g) - \widetilde{\theta}\|_2^2 = \big\|(M_n - I_p)\widetilde{\Sigma}_\theta^{1/2}\big\|_F^2 + \sum_{j=1}^{n} \mathrm{tr}(K_{n,j}\Gamma K_{n,j}^\top), \tag{52}$$

$$\sum_{n=1}^{T_\rho} \|\widetilde{x}_{n+1}^\top \widetilde{\theta}_{K+n}(g) - g^{\mathrm{Bayes}}_{K+n}\|_2^2 = \sum_{n=1}^{T_\rho} \mathbb{E}\|\widetilde{\theta}_{K+n}(g) - \widetilde{\theta}\|_2^2 - \mathcal{E}^{\mathrm{Bayes}}_{T_\rho}. \tag{53}$$

Scalar prediction norms denote L^2 norms. Both identities hold at finite horizons; the first gives estimation error, and prediction loss adds σ^2.

Proof. Write $r_j = \widetilde{\theta} + e_j$, where

$$e_j = (\widetilde{x}_j \widetilde{x}_j^\top - I_p)\widetilde{\theta} + \widetilde{x}_j \epsilon_{K+j}.$$

Conditionally on $\widetilde{\theta}$, the e_j are centered and independent. Gaussian fourth moments give

$$\mathbb{E}[\widetilde{\theta} e_j^\top] = 0, \qquad \mathbb{E}[e_j e_k^\top] = 0 \quad (j \ne k), \qquad \mathbb{E}[e_j e_j^\top] = \widetilde{\Sigma}_\theta + (\mathrm{tr}\,\widetilde{\Sigma}_\theta + \sigma^2) I_p = \Gamma.$$

Expanding $\widetilde{\theta}_{K+n}(g) - \widetilde{\theta} = (M_n - I_p)\widetilde{\theta} + \sum_{j\le n} K_{n,j} e_j$ proves (52). The next regressor $\widetilde{x}_{n+1}$ is fresh and isotropic, so coefficient error and prediction error are isometric. Posterior Pythagoras then subtracts $\mathbb{E}\,\mathrm{tr}\,P_n$ on row n; summing proves (53). □

F.2.1 Proof of Corollary 1: assembly floors

We first show that one attention layer can assemble only the current evidence atom (Lemma 15); Corollary 2 then gives its regret floor. For one-layer fixed-transition SSMs, Lemma 17 shows that convolution width $d_{\mathrm{conv}} \le K$ limits evidence assembly to at most d_{conv} atoms. Lemma 16 turns this limitation into a belief-alignment gap, which yields linear regret at fixed K through Theorem 2.

Lemma 15 (One attention layer assembles at most the current evidence atom)**.** *Fix $K \geq 1$, ambient width $d_e \geq 1$, $d, H, L \geq 1$, and $c_t \in \{1, \tau_t\}$. For every $g \in \mathcal{F}_{\mathrm{att}}^{\mathrm{cal}}(1, d, H, L; c_t)$ and every matched step n, its implicit belief has the orthogonal decomposition*

$$\widetilde{\theta}_{K+n}(g) = C_n r_n + \eta_n, \qquad \eta_n \perp \widetilde{\theta}, \qquad \eta_n \perp r_n, \tag{54}$$

where $r_n = \widetilde{x}_n Y_n$ is the current evidence atom and C_n is a deterministic matrix. If $L < K + 1$, then $C_n = 0$. In particular, no evidence atom r_j with $j < n$ appears in the Bayes-relevant projection.

Proof. Use the same homogenization as in the proof of Lemma 5: put $\bar{z}_s := (1, z_s^\top)^\top$. After combining the affine embedding, query, key, value, output, residual, and prediction maps, there are deterministic parameters a, $(M_h, v_h)_{h=1}^H$, and relative positional-bias sequences $(p_u^{(h)})$ such that, exactly,

$$g_t = a^\top \bar{z}_t + \frac{1}{c_t} \sum_{h=1}^{H} \sum_{j \in J_t} \left[p_{t-j}^{(h)} + \bar{z}_t^\top M_h \bar{z}_j \right] (v_h^\top \bar{z}_j). \tag{55}$$

Hold out the two adjacent-token occurrences of the fresh regressor $\widetilde{x}^\star := \widetilde{x}_{t+1-K}$. In one summand of (55), let $m_\star$ be the number of its internal affine factors from which $\widetilde{x}^\star$ is selected. Multiplication by the external $\widetilde{x}^\star$ contributes one further factor, so Gaussian parity eliminates even $m_\star$. If $K > 1$, the current token contains no $\widetilde{x}^\star$ and one source token contributes at most its key and value occurrences; hence $m_\star \leq 2$ and only $m_\star = 1$ survives. If $K = 1$, the current token is also a fresh-regressor token, so $m_\star \leq 3$; the additional surviving case $m_\star = 3$ is a deterministic Gaussian fourth-moment term. In either case the contribution to the implicit belief has degree at most two in the remaining design and response coordinates.

Project this contribution onto the degree-two space spanned by the evidence atoms $\{\widetilde{x}_{s-K} y_s : K < s \leq t\}$. A historical label y_s, $s < t$, occurs only in its own source token, while its routed regressor $\widetilde{x}_{s-K}$ occurs in two strictly earlier tokens. No summand of (55) contains all three of that label, its routed regressor, and the held-out regressor. Hence its kernel weight in the projection is zero by the independent Gaussian sign flips of the missing regressor. For $s = t$, however, y_t is in the current token and the source z_{t-K} contains both $\widetilde{x}_{t-K}$ and one occurrence of $\widetilde{x}^\star$; this is the unique routed degree-two contribution. It is available exactly when $t - K$ lies in the window, namely when $L \geq K + 1$. Collecting its deterministic kernel weights over all heads gives $C_n r_n$. By the degree-two projection argument of Lemma 4, the remainder η_n is orthogonal to both $\widetilde{\theta}$ and r_n. □

This leaves attention with only one noisy evidence atom for estimating the belief, yielding the following floor.

Corollary 2 (One-layer attention floor)**.** *For $c_t \in \{1, \tau_t\}$, every $g \in \mathcal{F}_{\mathrm{att}}^{\mathrm{cal}}(1, d, H, L; c_t)$ satisfies*

$$\mathrm{Reg}_T(g) \geq T_\rho \sum_{i=1}^{p} d_i - \mathcal{E}_{T_\rho}^{\mathrm{Bayes}} = \Omega(T_\rho). \tag{56}$$

The bound is uniform in d_e, d, H, L and all learned parameters. If $L < K + 1$, the coefficient $\sum_i d_i$ may be replaced by Λ.

Proof. Since $\widetilde{\theta}_{K+n}(g)$ is $\mathcal{G}_{K+n}^-$-measurable and the additional unpaired regressors leave the Bayes posterior unchanged, posterior Pythagoras gives

$$\mathbb{E}\|\widetilde{\theta}_{K+n}(g) - \widetilde{\theta}_{K+n}^{\mathrm{Bayes}}\|^2 = \mathbb{E}\|\widetilde{\theta}_{K+n}(g) - \widetilde{\theta}\|^2 - \mathbb{E}\,\mathrm{tr}\, P_n.$$

Lemma 15 and orthogonality let us discard η_n. The best linear estimate of $\widetilde{\theta}$ from the single evidence atom r_n has error covariance

$$\widetilde{\Sigma}_\theta - \widetilde{\Sigma}_\theta (\widetilde{\Sigma}_\theta + \Gamma)^{-1} \widetilde{\Sigma}_\theta,$$

whose trace is $\sum_i d_i$. Summing and using the prediction identity (50) proves the claim. If $C_n = 0$, the corresponding prior error is $\mathrm{tr}\, \widetilde{\Sigma}_\theta = \Lambda$. □

We now turn to SSMs, where a short convolution limits assembly to a bounded number of recent labels. We use two consequences of the general routing setup in §F.1: first, a persistent belief-alignment gap forces linear regret; then, limited evidence assembly creates such a gap.

After N_t^ρ matched observations, posterior Pythagoras gives

$$\|\widetilde{\theta}_t^{\mathrm{Bayes}}\|_t^2 = \Lambda - \mathbb{E}\,\mathrm{tr}\,P_{N_t^\rho}. \tag{57}$$

For $n > p+1$,

$$\mathbb{E}\,\mathrm{tr}\,P_n \le \sigma^2 \mathbb{E}\,\mathrm{tr}\left(\sum_{j=1}^{n} \widetilde{x}_j \widetilde{x}_j^\top\right)^{-1} = \frac{p\sigma^2}{n-p-1}.$$

Thus the Bayes belief energy approaches Λ and is at least $\Lambda/2$ after a finite $n_0 = n_0(p, \widetilde{\Sigma}_\theta, \sigma)$ matched observations. A uniform defect $\delta > 0$ in the product of the two factors in Theorem 2 on a positive fraction of matched targets therefore implies $\mathrm{Reg}_T(g) = \Omega(T_\rho)$; only the first n_0 matched targets are lost.

It remains to show why limited evidence assembly leaves a belief-alignment gap. Unlike the deterministic evidence-atom projection in (54), we now condition on the regressors and project $\widetilde{\theta}_t(g)$ onto the span of the matched labels:

$$\widetilde{\theta}_t(g) = \sum_{s \in \mathcal{M}_t} w_{t,s} y_s + \eta_t, \tag{58}$$

where $w_{t,s} \in \mathbb{R}^p$ is design-measurable and η_t is conditionally orthogonal to every matched label. The conditional label covariance is strictly positive definite because $\sigma^2 > 0$, so these coefficients are unique. The next lemma shows that full belief alignment is impossible if these coefficients correlate with their routed regressors for only a bounded number of labels.

Lemma 16 (Belief alignment under limited evidence assembly)**.** *Fix p, $C < \infty$, $\widetilde{\Sigma}_\theta \succ 0$, and $\sigma^2 > 0$. There are constants $n_C < \infty$ and $\delta_{\widetilde{\Sigma}_\theta, C} > 0$ such that the following holds. If the coefficients in (58) satisfy*

$$\mathbb{E}[w_{t,s} \widetilde{x}_{\rho(s)}^\top] = 0 \qquad (s \in \mathcal{M}_t \setminus \mathcal{J}_t) \tag{59}$$

for a deterministic set $\mathcal{J}_t \subseteq \mathcal{M}_t$ with $|\mathcal{J}_t| \le C$, then

$$\mathrm{corr}^2(\widetilde{\theta}_t(g), \widetilde{\theta}_t^{\mathrm{Bayes}}) \le 1 - \delta_{\widetilde{\Sigma}_\theta, C}$$

whenever $N_t^\rho \ge n_C$. The conclusion is uniform over all other model parameters for which the same bound on $|\mathcal{J}_t|$ holds.

Proof. The tower property and conditional posterior linearity give

$$\langle \widetilde{\theta}_t(g), \widetilde{\theta}_t^{\mathrm{Bayes}} \rangle_t = \mathbb{E}[\widetilde{\theta}_t(g)^\top \widetilde{\theta}] = \sum_{s \in \mathcal{M}_t} \mathbb{E}[w_{t,s}^\top \widetilde{\Sigma}_\theta \widetilde{x}_{\rho(s)}].$$

Put $\mathsf{C}_t := \sum_{s \in \mathcal{M}_t} w_{t,s} \widetilde{x}_{\rho(s)}^\top$ and denote the last display by m. Conditional orthogonality in (58) gives

$$\|\widetilde{\theta}_t(g)\|_t^2 \ge \mathbb{E}\,\mathrm{tr}(\mathsf{C}_t \widetilde{\Sigma}_\theta \mathsf{C}_t^\top) + \sigma^2 \mathbb{E} \sum_{s \in \mathcal{M}_t} \|w_{t,s}\|_2^2. \tag{60}$$

By matrix Cauchy–Schwarz,

$$\mathbb{E}\,\mathrm{tr}(\mathsf{C}_t \widetilde{\Sigma}_\theta \mathsf{C}_t^\top) \ge \frac{m^2}{\Lambda}.$$

Condition (59) reduces m to the terms in $\mathcal{J}_t$, and another Cauchy–Schwarz inequality gives

$$m^2 \le C\,\mathrm{tr}(\widetilde{\Sigma}_\theta^2) \mathbb{E} \sum_{s \in \mathcal{J}_t} \|w_{t,s}\|_2^2.$$

For $C > 0$, set $a_{\widetilde{\Sigma}_\theta, C} := \sigma^2 \Lambda / [C\,\mathrm{tr}(\widetilde{\Sigma}_\theta^2)]$. Then

$$\|\widetilde{\theta}_t(g)\|_t^2 \ge \frac{m^2}{\Lambda}(1 + a_{\widetilde{\Sigma}_\theta, C}).$$

By (57) and the posterior-variance bound following it, choose n_C so that

$$\|\widetilde{\theta}_t^{\mathrm{Bayes}}\|_t^2 \geq \Lambda \frac{2 + a_{\widetilde{\Sigma}_\theta, C}}{2(1 + a_{\widetilde{\Sigma}_\theta, C})}.$$

Then

$$\mathrm{corr}^2(\widetilde{\theta}_t(g), \widetilde{\theta}_t^{\mathrm{Bayes}}) \leq \frac{2}{2 + a_{\widetilde{\Sigma}_\theta, C}} = 1 - \frac{a_{\widetilde{\Sigma}_\theta, C}}{2 + a_{\widetilde{\Sigma}_\theta, C}}.$$

Take $\delta_{\widetilde{\Sigma}_\theta, C} := a_{\widetilde{\Sigma}_\theta, C}/(2 + a_{\widetilde{\Sigma}_\theta, C})$. If $C = 0$, then $\mathcal{J}_t = \varnothing$, so $m = 0$ and the squared correlation is zero directly. □

We now verify condition (59) for a one-layer fixed-transition SSM, including its count-normalized extension.

Lemma 17 (Limited evidence assembly in short-convolution SSMs)**.** *Fix $K, d_e, d, H \geq 1$, $1 \leq q \leq K$, and $c_t \in \{1, t\}$. Every $g \in \mathcal{F}_{\mathrm{SSM}}^{\mathrm{cal}}(1, d, H, q; c_t)$ satisfies the condition of Lemma 16 with*

$$\mathcal{J}_t = \{t - q + 1, \ldots, t\} \cap \{K + 1, \ldots, t\}.$$

Thus $|\mathcal{J}_t| \leq q$, and

$$\mathrm{Reg}_T\big(\mathcal{F}_{\mathrm{SSM}}^{\mathrm{cal}}(1, d, H, q; c_t)\big) = \Omega(T_\rho), \tag{61}$$

uniformly over d_e, d, H, $1 \leq q \leq K$, all learned convolution taps, and all fixed transition matrices.

Proof. Treat one head and unroll its fixed-transition state. Its contribution at time t is a sum over write times of a current affine read multiplied by a bilinear write; the transition changes only their deterministic coefficients. A width-q convolved token at time u can contain labels from positions $u - q + 1, \ldots, u$ and regressors from positions $u - q + 1, \ldots, u + 1$, after clipping to available indices.

Hold out the fresh routed regressor $\widetilde{x}^\star := \widetilde{x}_{t+1-K}$. After multiplying by $\widetilde{x}^\star$ and integrating it out to form the implicit belief, Gaussian parity leaves two possible sources of an evidence atom $\widetilde{x}_{s-K} y_s$. First, if the read supplies $\widetilde{x}^\star$, both fields of the atom must come from one write at some time r. If that write contains y_s, then

$$s - K \leq r - K \leq r - q,$$

whereas its earliest available regressor is $\widetilde{x}_{r-q+1}$, so it cannot also contain $\widetilde{x}_{s-K}$. Second, if a write supplies $\widetilde{x}^\star$, one remaining field comes from that write and one from the current read. If the read supplies y_s, then $s \in \{t - q + 1, \ldots, t\}$. If instead the write supplies y_s, the read would have to supply $\widetilde{x}_{s-K}$; this is impossible because

$$s - K \leq t - K < t - q + 1.$$

The fourth-moment terms and direct residual contribution are deterministic and contain no evidence atom.

The implicit belief has degree at most two. Conditional on the design, its response-even part is orthogonal to the matched labels, while its response-odd part is linear in those labels with affine design loadings. Thus $\mathbb{E}[w_{t,s} \widetilde{x}_{s-K}^\top]$ is the coefficient of the matched monomial $\widetilde{x}_{s-K} y_s$, and it vanishes outside $\mathcal{J}_t$, giving (59). Summing heads preserves the same set $\mathcal{J}_t$, and the deterministic factor $1/c_t$ does not change its support. Since $|\mathcal{J}_t| \leq q \leq K$, Lemma 16, the Bayes energy bound following (57), and Theorem 2 prove (61). □

Taking $c_t \equiv 1$ and $q = d_{\mathrm{conv}}$, together with Corollary 2, proves Corollary 1.

F.2.2 Constructions and upper bounds for Proposition 2

The constructions first assemble evidence and address the required regressor, then apply the Case 1 kernels for belief maintenance. Proposition 8 implements this with two attention layers and positional bias; Proposition 9 uses a one-layer SSM with a width-$(K + 1)$ convolution. We compute their regret using Lemma 14.

For the attention construction, we reuse the Case 1 window notation from Lemma 6 with horizon T_ρ and window $1 \le L_0 \le T_\rho$. With n indexing matched observations, write

$$\tau_n := \min(n, L_0), \qquad N_j := \sum_{n=1}^{T_\rho} \tau_n^j, \quad j = 1, 2.$$

The optimal uniform kernel's cumulative belief-estimation error is then $\mathcal{E}_{L_0,T_\rho}$ from (27).

Proposition 8 (Two-layer attention construction)**.** *Let $L \ge K+1$ and put $L_0 := \min(L, T_\rho)$. If the ambient width satisfies $d_e \ge 4p+2$, then for every $d, H \ge 1$ with $Hd \ge 2p$, there is a two-layer linear-attention predictor in $\mathcal{F}_{\mathrm{att}}(2, d, H, L)$ whose regret is*

$$\mathcal{E}_{L_0,T_\rho} - \mathcal{E}_{T_\rho}^{\mathrm{Bayes}}. \tag{62}$$

Let g_{box} denote the construction that weights every retained evidence atom by $1/L$. When $L \le T_\rho$, its regret is

$$\mathrm{Reg}_T(g_{\mathrm{box}}) = \Lambda\left(\frac{L}{3} - \frac{1}{2} + \frac{1}{6L}\right) + V\left(\frac{T_\rho}{L} - \frac{1}{2} + \frac{1}{2L}\right) - \mathcal{E}_{T_\rho}^{\mathrm{Bayes}}. \tag{63}$$

When $L > T_\rho$, the same predictor satisfies

$$\mathrm{Reg}_T(g_{\mathrm{box}}) = \Lambda T_\rho\left(1 - \frac{T_\rho+1}{L}\right) + \frac{T_\rho(T_\rho+1)}{6L^2}\big[\Lambda(2T_\rho+1) + 3V\big] - \mathcal{E}_{T_\rho}^{\mathrm{Bayes}}. \tag{64}$$

Proof. Use four p-dimensional residual blocks and two scalar coordinates, setting any remaining coordinates to zero. The initial embedding places the two whitened regressors and the label in the first two vector blocks and one scalar coordinate; the other two vector blocks and the accumulator coordinate are reserved for the two attention layers.

In layer one, set every data-dependent score to zero and spike the relative positional bias at offset K. Partition the $2p$ coordinates of the two source vectors among the H heads, with at most d coordinates assigned to each head. This is possible because $Hd \ge 2p$. The value map of each head extracts its assigned coordinates from z_{s-K}, and its output map writes them into the corresponding reserved residual coordinates at position s. Summing the head outputs therefore copies both $\widetilde{x}_{s+1-K}$ and $\widetilde{x}_{s-K}$ exactly. Thus the first layer supplies the next routed regressor and the regressor paired with y_s.

For layer two, partition the p eigencoordinates among the heads, again with at most d coordinates per head. On the coordinates assigned to head b, choose its query and key maps so that its score is

$$\sum_{i \in I_b} \gamma_i (U^\top \widetilde{x}_{n+1})_i (U^\top \widetilde{x}_j)_i.$$

Its value map reads Y_j into one head-output coordinate, and its output map writes the resulting scalar into the shared accumulator coordinate. Summing over heads gives

$$g_{K+n} = \widetilde{x}_{n+1}^\top U \,\mathrm{diag}(\gamma_1, \dots, \gamma_p) U^\top \sum_{j=n-\tau_n+1}^{n} r_j.$$

The final linear readout selects this scalar accumulator. The residual-width condition accounts for the original two vector blocks, the two routed blocks, the label, and the accumulator, while the headwise partitions respect the per-head query, key, value, and pre-projection output width d.

In the eigenbasis U, Lemma 14 separates the objective over coordinates. For coordinate i it is

$$\lambda_i T_\rho - 2\lambda_i \gamma_i N_1 + \gamma_i^2 (\lambda_i N_2 + v_i N_1).$$

Optimizing γ_i proves (62). If $L \le T_\rho$, taking $\gamma_i = 1/L$ and using

$$\sum_{n=1}^{T_\rho} (1 - \tau_n/L)^2 = \frac{L}{3} - \frac{1}{2} + \frac{1}{6L}, \qquad \sum_{n=1}^{T_\rho} \tau_n/L^2 = \frac{T_\rho}{L} - \frac{1}{2} + \frac{1}{2L}$$

gives (63). If $L > T_\rho$, then $\tau_n = n$, and the two sums are

$$\sum_{n=1}^{T_\rho} (1 - n/L)^2 = T_\rho - \frac{T_\rho(T_\rho+1)}{L} + \frac{T_\rho(T_\rho+1)(2T_\rho+1)}{6L^2}, \qquad \sum_{n=1}^{T_\rho} \frac{n}{L^2} = \frac{T_\rho(T_\rho+1)}{2L^2}.$$

Substitution proves (64). □

Applying Proposition 8 with the following window choice gives the attention upper bound in Proposition 2. Section F.2.3 supplies the matching lower bound when $K = O(\sqrt{T_\rho})$.

Recall from (16)– (17) the scales $\zeta, \bar{\zeta}$ and the nearest-integer common window $L^\star(T_\rho)$. To ensure positional reach, if $K + 1 \le T_\rho$, choose

$$L_0 = \min\{T_\rho, \max\{K + 1, L^\star(T_\rho)\}\}, \qquad L = L_0.$$

When $K + 1 \le L^\star(T_\rho)$ and $L^\star(T_\rho) = o(T_\rho)$,

$$\mathrm{Reg}_T(g_{\mathrm{box}}) = \frac{2}{\sqrt{3}}\bar{\zeta}\sqrt{T_\rho} + o(\sqrt{T_\rho}).$$

If $L^\star(T_\rho) < K + 1 \le T_\rho$, the same construction has

$$\mathrm{Reg}_T(g_{\mathrm{box}}) \le \frac{\Lambda(K+1)}{3} + \frac{VT_\rho}{K+1} - \mathcal{E}^{\mathrm{Bayes}}_{T_\rho} + O(1).$$

Finally, if $K + 1 > T_\rho$, choose $L = K + 1$. Equation (64) gives $O(K)$ regret, completing the attention upper bound in Proposition 2.

Proposition 9 (One-layer SSM construction)**.** *If $Hd \ge p$ and the ambient width satisfies $d_e \ge 2p+2$, the one-layer class $\mathcal{F}_{\mathrm{SSM}}(1, d, H, K+1)$ contains a predictor g_{exp} satisfying*

$$\mathrm{Reg}_T(g_{\mathrm{exp}}) = \zeta\sqrt{T_\rho} + o(\sqrt{T_\rho}). \tag{65}$$

The remainder is for fixed $p, \widetilde{\Sigma}_\theta, \sigma^2$.

Proof. The initial embedding exposes the two whitened-regressor blocks and the label in $2p + 1$ residual coordinates, leaving one clean output coordinate. At time $s = K + j$, the width-$(K + 1)$ coordinatewise convolution reads the next-regressor block of z_{s-K} as $\widetilde{x}_{j+1}$, its current-regressor block as $\widetilde{x}_j$, and the current label as $y_s = Y_j$. Distribute the p eigencoordinates among the Hd available state coordinates. In direction i, the bilinear write forms $(1 - \alpha_i)(U^\top \widetilde{x}_j)_i Y_j$, the fixed transition uses a scalar decay α_i, and the ordinary token–state read multiplies the state by $(U^\top \widetilde{x}_{j+1})_i$. This gives the whitened kernel

$$K_u = U\,\mathrm{diag}(\kappa_{u,1}, \dots, \kappa_{u,p})U^\top,$$
$$\kappa_{u,i} = (1 - \alpha_i)\alpha_i^u, \qquad \frac{1 + \alpha_i}{1 - \alpha_i} = 2\sqrt{\frac{v_i T_\rho}{\lambda_i}}.$$

Lemma 14 and the geometric-sum calculation of Lemma 3, applied coordinatewise, give $\sqrt{\lambda_i v_i T_\rho} + o(\sqrt{T_\rho})$ before the common $\mathcal{E}^{\mathrm{Bayes}}_{T_\rho} = O(\log T_\rho)$ subtraction. Summing the fixed number of coordinates proves the claim. □

This proves the SSM upper bound in Proposition 2. Section F.2.4 supplies the matching lower bound and shows that $Hd \ge p$ is necessary for sublinear regret throughout the one-layer class, regardless of convolution width.

F.2.3 Attention lower bounds for Propositions 2 and 3

Lemma 18 derives a regret lower bound from the limited number of matched observations available to fixed-depth attention. It proves Proposition 3 and, together with Proposition 8, the attention claims in Proposition 2. The argument depends only on which raw tokens can influence a prediction, so it applies to both linear and softmax attention and their count-normalized extensions.

Recall the Bayes estimation error from (8):

$$e_n^{\mathrm{Bayes}} = \mathbb{E}\,\mathrm{tr}\left(\widetilde{\Sigma}_\theta^{-1} + \sigma^{-2}\sum_{j=1}^{n}\widetilde{x}_j\widetilde{x}_j^\top\right)^{-1}, \tag{66}$$

with an empty sum at $n = 0$. This is the posterior error after n matched observations.

Lemma 18 (Finite-context attention floor)**.** *Fix $1 \le K < T$ and recall $T_\rho = T - K$. Across D layers, a prediction depends on at most $L_D := 1 + D(L-1)$ raw tokens. For every ambient width $d_e \ge 1$, every $D, d, H, L \ge 1$, every admissible count-normalization schedule $\mathbf{c}$, and every $g \in \mathcal{F}_{\mathrm{att}}^{\mathrm{sm,cal}}(D, d, H, L; \mathbf{c})$,*

$$\mathrm{Reg}_T(g) \ge \sum_{j=L_D+1}^{T_\rho} \left[e_{L_D}^{\mathrm{Bayes}} - e_j^{\mathrm{Bayes}} \right]. \tag{67}$$

If $T_\rho \ge 4(L_D + \sigma^2/\min_i \lambda_i + p + 2)$, then

$$\mathrm{Reg}_T(g) \ge \frac{p\sigma^2 T_\rho}{4(L_D + \sigma^2/\min_i \lambda_i)} \ge \frac{p\sigma^2}{4(1 + \sigma^2/\min_i \lambda_i)} \frac{T_\rho}{L_D}. \tag{68}$$

Both bounds are uniform in d_e, d, H and all attention parameters. Taking $\mathbf{c} = \mathbf{1}$ gives the same statement for the softmax-enabled class $\mathcal{F}_{\mathrm{att}}^{\mathrm{sm}}(D, d, H, L)$, and hence for its all-linear subclass $\mathcal{F}_{\mathrm{att}}(D, d, H, L)$. Consequently, any fixed-depth attention sequence with $O(\sqrt{T_\rho})$ regret must have $L = \Omega(\sqrt{T_\rho})$.

Proof. Let $I_{r,t} := [\max\{1, t - r(L-1)\}, t]$. Induction over layers shows that the representation at position t after layer r is measurable with respect to $\sigma(z_j : j \in I_{r,t})$. Indeed, layer r accesses only the L layer- $(r-1)$ representations ending at t, and the union of their raw intervals is $I_{r,t}$. Scores, either attention operator, deterministic count normalization, values, residual updates, and multi-head recombination are measurable functions of those representations. Hence the final output depends on at most L_D raw token positions. Cached evaluation does not enlarge this computation graph: a cached key or value is frozen when its position is computed.

Consider matched step n, corresponding to prediction time $t = K + n$. Enlarge the visible field by revealing every regressor in the entire history, while retaining only the labels from $I_{D,t}$. This genie field contains all variables on which the attention output depends. At most $\min(n, L_D)$ of its labels are matched, and their routed regressors are distinct standard Gaussian vectors. All other revealed regressors are independent of $\widetilde{\theta}$ and contribute no likelihood factor. Gaussian conjugacy therefore makes the minimum coefficient estimation error under the genie field $e_{\min(n,L_D)}^{\mathrm{Bayes}}$, whereas the full-history Bayes estimation error is e_n^{Bayes}. The next routed regressor is fresh and isotropic, so this difference lower-bounds the excess prediction loss. Conditional-expectation Pythagoras gives

$$\ell_{K+n}(g) - \ell_{K+n}(g^{\mathrm{Bayes}}) \ge e_{\min(n,L_D)}^{\mathrm{Bayes}} - e_n^{\mathrm{Bayes}}.$$

The terms with $n \le L_D$ vanish. Summing the remaining rows proves (67).

For the explicit rate, operator Jensen and the Wishart inverse moment give

$$e_{L_D}^{\mathrm{Bayes}} \ge \mathrm{tr}(\widetilde{\Sigma}_\theta^{-1} + L_D \sigma^{-2} I_p)^{-1} \ge \frac{p\sigma^2}{L_D + \sigma^2/\min_i \lambda_i}, \tag{69}$$

$$e_j^{\mathrm{Bayes}} \le \sigma^2 \mathbb{E}\,\mathrm{tr}\left(\sum_{i=1}^{j} \widetilde{x}_i \widetilde{x}_i^\top \right)^{-1} = \frac{p\sigma^2}{j - p - 1}, \qquad j > p + 1. \tag{70}$$

Thus every $j \ge \lceil 2(L_D + \sigma^2/\min_i \lambda_i) + p + 1 \rceil$ contributes at least $p\sigma^2/[2(L_D + \sigma^2/\min_i \lambda_i)]$ to (67). The stated condition on T_ρ leaves at least $T_\rho/2$ such indices, proving the first inequality in (68); the second uses $L_D + \sigma^2/\min_i \lambda_i \le (1 + \sigma^2/\min_i \lambda_i) L_D$. Finally, if the displayed condition fails, then $L_D = \Omega(T_\rho)$ already. If it holds, the clean bound and $O(\sqrt{T_\rho})$ regret imply $L_D = \Omega(\sqrt{T_\rho})$. At fixed D, $L_D = 1 + D(L-1) = \Theta(L)$, which proves the context requirement. □

Taking $\mathbf{c} = \mathbf{1}$ proves Proposition 3. When $K = O(\sqrt{T_\rho})$, the window $L = \widetilde{L}^\star$ in Proposition 2 satisfies $L = \Theta(\sqrt{T_\rho})$. At depth two, (68) therefore gives $\Omega(\sqrt{T_\rho})$ regret. Together with the upper bound in §F.2.2, this completes the attention claims in Proposition 2.

F.2.4 SSM LOWER BOUNDS FOR PROPOSITION 2

The construction in Proposition 9 gives the upper bound. For the lower bounds, Lemma 19 reduces the implicit belief to a stationary kernel with a boundary term for the first atom. Lemma 20 proves linear regret when $Hd < p$, at any convolution width. Proposition 10 combines these lemmas with the scalar stationary-kernel bound of Lemma 3 to prove both SSM claims in Proposition 2. We conclude with the memory comparison.

Lemma 19 (SSM evidence projection with a boundary term)**.** *Fix $K, d_e, d, H \geq 1$, and let*

$$g \in \mathcal{F}_{\mathrm{SSM}}(1, d, H, K{+}1)$$

use arbitrary learned head-specific convolution taps and arbitrary fixed transition matrices. There are deterministic matrices $W_{\mathrm{read}} \in \mathbb{R}^{p\times Hd}$, $\Phi \in \mathbb{R}^{Hd\times Hd}$, and $W_{\mathrm{write}} \in \mathbb{R}^{Hd\times p}$, together with kernel corrections $\Delta K_0, \dots, \Delta K_K, \Delta K_0^{\partial}, \dots, \Delta K_K^{\partial} \in \mathbb{R}^{p\times p}$, such that at matched step n its implicit belief satisfies

$$\begin{aligned}\widetilde{\theta}_{K+n}(g) = \sum_{j=1}^{n} W_{\mathrm{read}}\Phi^{n-j}W_{\mathrm{write}}r_j + \sum_{j=\max\{2,n-K\}}^{n} \Delta K_{n-j}r_j \\ + \mathbf{1}_{\{n\leq K+1\}}\Delta K_{n-1}^{\partial}r_1 + \eta_n,\end{aligned} \tag{71}$$

where $\mathbb{E}[\eta_n\widetilde{\theta}^{\top}] = 0$ and $\mathbb{E}[\eta_n r_j^{\top}] = 0$ for $1 \leq j \leq n$. Thus every routed kernel weight at lag at least $K+1$ has range in the common subspace $\mathrm{col}(W_{\mathrm{read}})$, whose dimension is at most Hd. Equivalently, the same belief has the coarser decomposition

$$\widetilde{\theta}_{K+n}(g) = A_n r_1 + \sum_{j=2}^{n} K_{n-j}r_j + \eta_n, \tag{72}$$

where $A_n \in \mathbb{R}^{p\times p}$ may vary with n, whereas the matrices $K_u \in \mathbb{R}^{p\times p}$ depend only on the lag $u = n - j$. The assertion is uniform in the ambient and state dimensions, parameter magnitudes, and spectra of the fixed transitions.

Proof. It is enough to analyze one head and then sum the resulting projections. Absorb the fixed whitening $x \mapsto \Sigma_x^{-1/2}x$ into the affine maps. For head b,

$$\widetilde{e}_s^{(b)} = \sum_{j=0}^{K} C_j^{(b)} \odot e_{s-j}^{(0)}, \qquad u_s^{(b)} = (W_K^{(b)}\widetilde{e}_s^{(b)}) \odot (W_V^{(b)}\widetilde{e}_s^{(b)}), \qquad h_t^{(b)} = \sum_{s\leq t}(\Phi^{(b)})^{t-s}u_s^{(b)}.$$

The slice of the final linear map acting on this head turns its token–state read into $q_t^{(b)\top}h_t^{(b)}$, where $q_t^{(b)}$ is affine in $\widetilde{e}_t^{(b)}$. Biases in the affine factors are included by homogenizing with a constant coordinate.

Computing the implicit belief. No observed label uses the fresh regressor $\widetilde{x}_{n+1}$. After holding out its two adjacent-token occurrences, expand, for $i, j, k \in \{1, \dots, p\}$,

$$\begin{aligned} q_t^{(b)} &= q_{t,0}^{(b)} + \sum_i \widetilde{x}_{n+1,i}q_{t,i}^{(b)}, \\ u_s^{(b)} &= u_{s,0}^{(b)} + \sum_j \widetilde{x}_{n+1,j}u_{s,j}^{(b)} + \sum_{j,k}\widetilde{x}_{n+1,j}\widetilde{x}_{n+1,k}u_{s,jk}^{(b)}. \end{aligned}$$

All displayed coefficients are free of $\widetilde{x}_{n+1}$. The vectors $q_{t,i}^{(b)}$ and $u_{s,jk}^{(b)}$ are deterministic: they are the coefficients of $\widetilde{x}_{n+1}$ in an affine read and of $\widetilde{x}_{n+1,j}\widetilde{x}_{n+1,k}$ in a product of two affine write factors. The remaining coefficients $q_{t,0}^{(b)}$ and $u_{s,j}^{(b)}$ have degree at most one in the other token fields, while $u_{s,0}^{(b)}$ has degree at most two.

Put $u := t - s$. Gaussian parity leaves only terms with one or three internal fresh-regressor factors, giving

$$\mathbb{E}\Big[\widetilde{x}_{n+1,i}q_t^{(b)\top}(\Phi^{(b)})^u u_s^{(b)} \mid \mathcal{G}_t^-\Big] = q_{t,i}^{(b)\top}(\Phi^{(b)})^u u_{s,0}^{(b)} + q_{t,0}^{(b)\top}(\Phi^{(b)})^u u_{s,i}^{(b)} + c_{i,u}^{(b)}, \tag{73}$$

where $c_{i,u}^{(b)}$ collects the three-factor terms and is deterministic because their coefficients $q_{t,i}^{(b)}$ and $u_{s,jk}^{(b)}$ are deterministic. Thus every nonconstant term on the right of (73) has total degree at most two in the remaining design and response fields.

The proof of Lemma 4 now applies to the matched pairs $(\widetilde{x}_j, Y_j)$, even though their two fields occupy different raw tokens. Partial injectivity makes the $\widetilde{x}_j$'s distinct. The paired flip $(\widetilde{x}_j, \epsilon_{K+j}) \mapsto (-\widetilde{x}_j, -\epsilon_{K+j})$ fixes every other routed evidence atom and shows that a cross-atom $\widetilde{x}_i Y_j$, $i \neq j$, is orthogonal to both $\widetilde{\theta}$ and the routed evidence space. The global response flip treats the zero- and two-response terms. Consequently the projection of (73) onto the routed evidence has deterministic matrix-valued kernel weights, and every remaining term belongs to a residual that is coordinatewise orthogonal to both $\widetilde{\theta}$ and that evidence. Equivalently, after summing these terms into η_n, $\mathbb{E}[\eta_n \widetilde{\theta}^\top] = 0$ and $\mathbb{E}[\eta_n r_j^\top] = 0$ for every $j \leq n$.

Regrouping by label time. At write time s, a width-$(K+1)$ convolution can see regressors from $s-K$ through $s+1$ and labels from $s-K$ through s, clipped to available indices. Fixed coordinate mixing changes the matrix coefficient of a monomial but not these time indices.

First consider the term in (73) in which the read supplies the fresh regressor and the $\widetilde{x}_{n+1}$-free write supplies a matched pair. A label at time $r \leq s$ has its regressor available only if $r - K \geq s - K$, forcing $r = s$. Hence the only possible direct routed write is $\widetilde{x}_{s-K} y_s$. For head b, let $W_{\mathrm{write}}^{(b)} \in \mathbb{R}^{d \times p}$ be its effective evidence-write map, and let $W_{\mathrm{read}}^{(b)} \in \mathbb{R}^{p \times d}$ collect the coefficients with which the read supplies the fresh regressor. The state-mediated kernel weight on r_j at routed lag $u = n - j$ is therefore

$$W_{\mathrm{read}}^{(b)} (\Phi^{(b)})^u W_{\mathrm{write}}^{(b)}.$$

The other nonconstant term in (73) takes one occurrence of $\widetilde{x}_{n+1}$ from a write at $s = t - u$ and one remaining field from the current read. Such a write exists only for $0 \leq u \leq K$. If the read supplies the label $y_r = y_{t-v}$, then $0 \leq v \leq K$, and the remaining write factor can be the routed regressor $\widetilde{x}_{t-v-K}$ exactly when

$$0 \leq v \leq u \leq K.$$

Indeed, its regressor tap is $K + v - u$; when both adjacent-token occurrences are available, the duplicate next-regressor path has tap $K + v - u + 1$. For fixed $v = t - r$, summing over these u therefore gives a fixed bypass matrix depending only on v. If instead the read supplies the routed regressor and the write supplies the label, the only possibility is $v = u = 0$, which adds another lag-zero bypass. Thus all such bypasses act on the $K+1$ most recent routed atoms. Denote their head-b matrices by $\Delta K_0^{(b)}, \ldots, \Delta K_K^{(b)}$.

The first routed regressor $\widetilde{x}_1$ lacks its duplicate next-regressor occurrence because $e_0^{(0)} = W_{\mathrm{emb}} z_0 = 0$. This can alter a bypass correction while r_1 is among the $K+1$ most recent atoms; denote its resulting bypass matrices by $\Delta K_0^{\partial,(b)}, \ldots, \Delta K_K^{\partial,(b)}$. Its direct routed write is unaffected, because it uses the current occurrence of $\widetilde{x}_1$ in z_1 at tap K. Consequently, once $n \geq K+2$, the kernel weight on r_1 is exactly $W_{\mathrm{read}}^{(b)} (\Phi^{(b)})^{n-1} W_{\mathrm{write}}^{(b)}$. These cases exhaust the nonconstant terms in (73); the fourth-moment term $c_{i,u}^{(b)}$ is deterministic and has zero projection onto routed evidence.

Finally, concatenate the heads:

$$W_{\mathrm{read}} := [W_{\mathrm{read}}^{(1)} \ \cdots \ W_{\mathrm{read}}^{(H)}], \qquad \Phi := \mathrm{diag}(\Phi^{(1)}, \ldots, \Phi^{(H)}), \qquad W_{\mathrm{write}} := \begin{bmatrix} W_{\mathrm{write}}^{(1)} \\ \vdots \\ W_{\mathrm{write}}^{(H)} \end{bmatrix}.$$

Summing the bypass matrices and residuals over heads proves (71). Defining

$$A_n := W_{\mathrm{read}} \Phi^{n-1} W_{\mathrm{write}} + \mathbf{1}_{\{n \leq K+1\}} \Delta K_{n-1}^{\partial}, \qquad K_u := W_{\mathrm{read}} \Phi^u W_{\mathrm{write}} + \mathbf{1}_{\{u \leq K\}} \Delta K_u$$

then gives the coarser stationary-profile form (72). □

The first evidence atom needs a separate kernel weight even for $p = d = H = 1$ and $K = 1$. Choose convolved residual coordinates

$$\widetilde{e}_{s,1} := (z_s)_1 + (z_{s-1})_1, \qquad \widetilde{e}_{s,2} := y_s,$$

write $u_s = \widetilde{e}^2_{s,1}$, propagate with a nonzero scalar decay, and read with $y_t h_t$. The lag-zero kernel weight on the first routed atom can then differ from the lag-zero kernel weight on every later routed atom because x_1, unlike later regressors, has no occurrence in a preceding token: position 0 is zero-padded.

The same calculation of the implicit belief exposes a state-dimension bottleneck that arbitrary convolution cannot remove: outside the at-most-Hd- dimensional current-read image, matching an additional belief direction through the remaining bilinear read–write term necessarily incurs label noise.

Lemma 20 (State-dimension floor at any convolution width)**.** *Fix p, $\widetilde{\Sigma}_\theta \succ 0$, and $\sigma^2 > 0$. For $1 \le K < T$ and $d_{\mathrm{conv}}, d_e, d, H \ge 1$, suppose $Hd < p$. Every $g \in \mathcal{F}_{\mathrm{SSM}}(1, d, H, d_{\mathrm{conv}})$ satisfies*

$$\mathrm{Reg}_T(g) \ge T_\rho \, \frac{\lambda_{\min}(\widetilde{\Sigma}_\theta)\sigma^2}{\sigma^2 + 2(p-1)\lambda_{\min}(\widetilde{\Sigma}_\theta)} - \mathcal{E}^{\mathrm{Bayes}}_{T_\rho}. \tag{74}$$

Thus the class has $\Omega(T_\rho)$ regret, uniformly over all remaining parameters, even when they depend on T.

Proof. Fix a matched step n, put $t = K + n$, and hold out the fresh regressor $\widetilde{x}_{n+1}$. Concatenate the head states into $h_t \in \mathbb{R}^{Hd}$ and write the scalar output as $g_t^{\mathrm{dir}} + q_t^\top h_t$, absorbing the output maps into the affine current read. The direct residual g_t^{dir} is affine, so its deterministic held-out-regressor coefficient is included in c_i below. The fresh-regressor calculation in the proof of Lemma 19, before its regrouping-by-label-time step, uses only the linearity of convolution and therefore applies at any convolution width. Thus

$$q_t = q_0 + A\widetilde{x}_{n+1}, \qquad h_t = h_0 + \sum_i \widetilde{x}_{n+1,i} h_i + \sum_{i,j} \widetilde{x}_{n+1,i}\widetilde{x}_{n+1,j} h_{ij},$$

where $A \in \mathbb{R}^{Hd \times p}$ and the h_{ij} are deterministic, q_0 and the h_i are affine in the remaining observed fields, and h_0 has degree at most two. The conditional expectation defining the implicit belief gives

$$\widetilde{\theta}_{K+n,i}(g) = A_{:i}^\top h_0 + q_0^\top h_i + c_i \tag{75}$$

for deterministic c_i.

Choose a possibly row-dependent unit vector $a \in \ker A$ and set $h_a := \sum_i a_i h_i$. Such a vector exists because $Hd < p$. Equation (75) gives

$$a^\top \widetilde{\theta}_{K+n}(g) = q_0^\top h_a + \text{constant}.$$

Let X stack each distinct remaining regressor once and put $Y := (Y_1, \dots, Y_n)^\top$. Since q_0 and h_a are affine, write

$$q_0 = b + BX + CY, \qquad h_a = e + EX + FY.$$

The response-odd mixed component of their product is $X^\top MY$, where

$$M = B^\top F + E^\top C, \qquad \mathrm{rank}(M) \le 2Hd.$$

The global response flip makes every response-even term orthogonal to both the target and the response-odd terms. Flipping all regressors and matched noises while holding $\widetilde{\theta}$ fixed then makes the remaining label-linear term odd, while $X^\top MY$ and $a^\top \widetilde{\theta}$ remain invariant. Hence

$$\mathbb{E}\big(a^\top \widetilde{\theta}_{K+n}(g) - a^\top \widetilde{\theta}\big)^2 \ge \mathbb{E}\big(X^\top MY - a^\top \widetilde{\theta}\big)^2. \tag{76}$$

Let $\mu \in \mathbb{R}^p$ be the sum of the matched blocks of M, pairing the block for $\widetilde{x}_j$ with the column for Y_j. Since $Y_j = \widetilde{x}_j^\top \widetilde{\theta} + \epsilon_{K+j}$, retaining the complete matrix M gives

$$\mathbb{E}\big(X^\top MY - a^\top \widetilde{\theta}\big)^2 \ge (\mu - a)^\top \widetilde{\Sigma}_\theta (\mu - a) + \sigma^2 \|M\|_F^2. \tag{77}$$

Indeed, the independent label noise contributes exactly $\sigma^2\|M\|_F^2$, while Jensen's inequality applied to the remaining quadratic-design term gives the first term.

We next use the rank bound on M to relate the matched coefficient μ to the noise cost. For $v \in \mathbb{R}^p$, define E_v with the same dimensions as M: column j contains v in the rows corresponding to $\widetilde{x}_j$ and zeros elsewhere. Its inner product with M therefore selects exactly the matched blocks, giving

$$\langle M, E_v \rangle_F = v^\top \mu.$$

Since the matched regressors are distinct, these columns are orthogonal, each with norm $\|v\|_2$. Hence $\|E_v\|_{\mathrm{op}} = \|v\|_2$. A singular-value decomposition and $\mathrm{rank}(M) \le 2Hd$ now give

$$|v^\top \mu| \le \sqrt{\mathrm{rank}(M)}\, \|M\|_F \|E_v\|_{\mathrm{op}} \le \sqrt{2Hd}\, \|M\|_F \|v\|_2, \qquad \|\mu\|_2^2 \le 2Hd \|M\|_F^2.$$

Thus the noise term is at least $\sigma^2 \|\mu\|_2^2/(2Hd)$. Combining this with (76)– (77) and completing the square gives

$$\begin{aligned}
\mathbb{E}\big(a^\top \widetilde{\theta}_{K+n}(g) - a^\top \widetilde{\theta}\big)^2 &\ge \inf_{\mu \in \mathbb{R}^p} \left\{ (\mu - a)^\top \widetilde{\Sigma}_\theta (\mu - a) + \frac{\sigma^2}{2Hd} \|\mu\|_2^2 \right\} \\
&= a^\top \left(\widetilde{\Sigma}_\theta^{-1} + \frac{2Hd}{\sigma^2} I_p \right)^{-1} a \\
&\ge \frac{\lambda_{\min}(\widetilde{\Sigma}_\theta)\sigma^2}{\sigma^2 + 2(p-1)\lambda_{\min}(\widetilde{\Sigma}_\theta)}.
\end{aligned}$$

The coefficient estimation error dominates this scalar projection on every matched step. Summing the routed-regressor and posterior Pythagorean identities proves (74). The same proof allows wider bilinear write features linearly combined into the state, and reads pairing affine token features with linear state projections. These preserve the quadratic state and affine q_t, so (75) and the rank bound $2Hd$ are unchanged, regardless of the intermediate feature widths. □

We now combine the preceding lemmas with the Case 1 stationary-kernel bound and the earlier construction to establish the SSM class regret.

Proposition 10 (SSM class regret)**.** *Fix p, $\widetilde{\Sigma}_\theta \succ 0$, and $\sigma^2 > 0$, and let $1 \le K < T$ with $T_\rho = T - K \to \infty$; K may depend on T. For every $d_e, d, H \ge 1$,*

$$\mathrm{Reg}_T\big(\mathcal{F}_{\mathrm{SSM}}(1, d, H, K{+}1)\big) \ge \sqrt{T_\rho - 1} \sum_{i=1}^{p} \sqrt{d_i v_i} - \mathcal{E}_{T_\rho}^{\mathrm{Bayes}} + o(\sqrt{T_\rho}). \tag{78}$$

If in addition $Hd \ge p$ and $d_e \ge 2p + 2$, then

$$\mathrm{Reg}_T\big(\mathcal{F}_{\mathrm{SSM}}(1, d, H, K{+}1)\big) \le \zeta \sqrt{T_\rho} + o(\sqrt{T_\rho}), \tag{79}$$

and hence the class regret is $\Theta(\sqrt{T_\rho})$. The bounds establish the rate but need not have sharp constants. If instead $Hd < p$, then for every $d_{\mathrm{conv}} \ge 1$,

$$\mathrm{Reg}_T\big(\mathcal{F}_{\mathrm{SSM}}(1, d, H, d_{\mathrm{conv}})\big) = \Omega(T_\rho). \tag{80}$$

Proof. Apply Lemma 19 and discard the orthogonal residual. Let $M_n := \sum_{u<n-1} K_u$. Before Bayes subtraction, the row estimation error involving the unrestricted kernel weight on the first atom is

$$\mathrm{tr}\Big[(A_n + M_n - I_p)\widetilde{\Sigma}_\theta (A_n + M_n - I_p)^\top\Big] + \mathrm{tr}(A_n \Gamma A_n^\top) + \sum_{u<n-1} \mathrm{tr}(K_u \Gamma K_u^\top).$$

Its exact matrix minimum over A_n is obtained at

$$A_n = -(M_n - I_p)\widetilde{\Sigma}_\theta (\widetilde{\Sigma}_\theta + \Gamma)^{-1}$$

and, using the eigenbasis U of $\widetilde{\Sigma}_\theta$, equals

$$\mathrm{tr}\big[(M_n - I_p) U D_d U^\top (M_n - I_p)^\top\big] + \sum_{u<n-1} \mathrm{tr}(K_u \Gamma K_u^\top), \qquad D_d := \mathrm{diag}(d_1, \dots, d_p).$$

For $n = 1$, $M_1 = 0$ and the minimized row is $\mathrm{tr}(D_d) \ge 0$, so we may discard it. Reindexing the remaining rows $n = 2, \dots, T_\rho$ by $n - 1$ gives the stationary-prefix functional at horizon $T_\rho - 1$. In

this basis, the bias and variance matrices are D_d and $U^\top \Gamma U = \mathrm{diag}(v_1, \ldots, v_p)$. Every off-diagonal entry of $U^\top K_u U$ therefore contributes only a nonnegative square. Discarding those entries leaves, in direction i, the stationary-prefix functional with bias coefficient d_i, variance coefficient v_i, and horizon $T_\rho - 1$. Lemma 3 yields $\sqrt{d_i v_i (T_\rho - 1)} + o(\sqrt{T_\rho})$, uniformly over that profile. Summing the fixed number of coordinates and subtracting the cumulative Bayes estimation error $\mathcal{E}_{T_\rho}^{\mathrm{Bayes}}$ proves (78). The construction in Proposition 9 lies in the class whenever $Hd \geq p$ and $d_e \geq 2p + 2$, and therefore proves (79). Finally, $\mathcal{E}_{T_\rho}^{\mathrm{Bayes}} = O(\log T_\rho) = o(\sqrt{T_\rho})$ and all d_i, v_i are positive, which gives the $\sqrt{T_\rho}$ rate. When $Hd < p$, Lemma 20 gives (80). □

This completes the SSM claims in Proposition 2. Convolution supplies positional reach, while the recurrent state must still span at least p dimensions. With $d_e \geq 2p + 2$, $d_{\mathrm{conv}} = K + 1$ and $Hd \geq p$ suffice for $\Theta(\sqrt{T_\rho})$ regret.

Memory comparison. For the SSM construction in Proposition 9, take $D = H = 1$, $d = p$, and $d_e = 2p + 2$. Its learned parameters, recurrent state, and convolution buffers together occupy $O(p^2 + p(K+1))$ scalar entries, hence $O(K+1)$ for fixed p. By Lemma 18, any fixed-depth attention achieving $O(\sqrt{T_\rho})$ regret requires $L = \Omega(\sqrt{T_\rho})$. With standard per-position caching of token representations or keys and values, this entails $\Omega(\sqrt{T_\rho})$ scalar entries. Thus when $K = o(\sqrt{T_\rho})$, the SSM attains the same regret rate with asymptotically less total storage.

F.3 Case 2.b: content routing

We first show how limited addressing capacity forces linear regret (§F.3.1), then develop geometric bounds on capacity (§F.3.2) and apply them to the architecture classes (§F.3.3). Section F.3.4 combines these ingredients to prove Theorem 3. Section F.3.5 then gives the attention and selective-SSM constructions for Proposition 4.

F.3.1 Addressing capacity and regret

We first establish how addressing capacity controls regret. To do so, we isolate each candidate regressor's component along the Bayes-belief direction.

We express the whitened variables from (3) in the eigenbasis U of the whitened prior covariance $\widetilde{\Sigma}_\theta$. Fix $p \in \mathbb{N}$, $\Sigma_x \succ 0$, $\Sigma_\theta \succ 0$, and $\sigma^2 > 0$, and write

$$\widetilde{\Sigma}_\theta := \Sigma_x^{1/2} \Sigma_\theta \Sigma_x^{1/2} = U \,\mathrm{diag}(\lambda_1, \ldots, \lambda_p) U^\top, \quad \xi_s := U^\top \widetilde{x}_s, \quad \beta := U^\top \widetilde{\theta}. \tag{81}$$

Thus $\xi_s \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, I_p)$, $\beta \sim \mathcal{N}(0, \mathrm{diag}(\lambda_1, \ldots, \lambda_p))$, and a routed label is $y_s = \xi_{\rho(s)}^\top \beta + \epsilon_s$. Fixed linear maps in either architecture implement the change of coordinates in (81).

Fix a prediction time $t \geq R$ in the content-routing model. Let $\mathcal{U}_t$ be the set of key indices not yet visited in the cycle containing $t + 1$, and set $N_t := |\mathcal{U}_t|$. The upcoming key index is uniform on $\mathcal{U}_t$. For $q \in \mathcal{U}_t$, let $X_q \in \mathbb{R}^p$ be the transformed regressor ξ_s at the most recent occurrence of key q.

To isolate the candidate vectors, we condition on the remaining history. Let $\mathcal{C}_t^-$ be the sigma-field generated by $z_{1:t}$ after removing the upcoming key ψ_{t+1} and every occurrence of the candidate vectors $(X_q)_{q \in \mathcal{U}_t}$. Conditional on $\mathcal{C}_t^-$,

$$(X_q)_{q \in \mathcal{U}_t} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, I_p), \tag{82}$$

independently of the upcoming key. The Bayes belief $\hat{\beta}_t^{\mathrm{Bayes}} := \mathbb{E}[\beta \mid z_{1:t}]$ is $\mathcal{C}_t^-$-measurable. The predictor may still depend arbitrarily on the candidate vectors through their earlier occurrences in the input. The arguments below preserve this dependence.

On $\{\|\hat{\beta}_t^{\mathrm{Bayes}}\| > 0\}$ define $\bar{\beta}_t^{\mathrm{Bayes}} := \hat{\beta}_t^{\mathrm{Bayes}} / \|\hat{\beta}_t^{\mathrm{Bayes}}\|$; when the norm vanishes, choose an arbitrary deterministic unit vector. Put

$$Z_q := (\bar{\beta}_t^{\mathrm{Bayes}})^\top X_q, \qquad X_q^\perp := (I_p - \bar{\beta}_t^{\mathrm{Bayes}} (\bar{\beta}_t^{\mathrm{Bayes}})^\top) X_q, \qquad \mathcal{C}_t := \mathcal{C}_t^- \vee \sigma(X_q^\perp : q \in \mathcal{U}_t). \tag{83}$$

Conditional on $\mathcal{C}_t$,

$$(Z_q)_{q\in\mathcal{U}_t} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad (\hat{\beta}_t^{\text{Bayes}})^\top X_q = \|\hat{\beta}_t^{\text{Bayes}}\| Z_q. \tag{84}$$

The second identity also holds when $\hat{\beta}_t^{\text{Bayes}} = 0$ because both sides then vanish. For $q \in \mathcal{U}_t$, let $g_{t,q}$ denote the predictor's output with the upcoming key fixed to q and all other input coordinates unchanged.

For conditionally square-integrable V, W, set

$$\text{corr}^2_{\mathcal{C}_t}(V,W) := \frac{\mathbb{E}[VW \mid \mathcal{C}_t]^2}{\mathbb{E}[V^2 \mid \mathcal{C}_t]\,\mathbb{E}[W^2 \mid \mathcal{C}_t]}.$$

The addressing capacity is

$$\text{Cap}_t(g) := \sum_{q\in\mathcal{U}_t} \text{corr}^2_{\mathcal{C}_t}(g_{t,q}, Z_q). \tag{85}$$

For $p = 1$, there are no orthogonal components to condition on.

For comparison with Theorem 2, fix $q \in \mathcal{U}_t$ and let

$$\mathcal{G}^-_{t,q} := \mathcal{C}_t \vee \sigma(Z_r : r \in \mathcal{U}_t,\ r \neq q), \qquad \hat{a}_{t,q}(g) := \mathbb{E}[Z_q g_{t,q} \mid \mathcal{G}^-_{t,q}].$$

Because Z_q is conditionally standard Gaussian and independent of $\mathcal{G}^-_{t,q}$, conditional orthogonal projection gives $g_{t,q} = \hat{a}_{t,q}(g) Z_q + \varepsilon_{t,q}$, where $\varepsilon_{t,q}$ is orthogonal to every cZ_q with $c \in L^2(\mathcal{G}^-_{t,q})$. Therefore,

$$\text{corr}^2_{\mathcal{C}_t}(g_{t,q}, \|\hat{\beta}_t^{\text{Bayes}}\| Z_q) = \underbrace{\text{corr}^2_{\mathcal{C}_t}(\hat{a}_{t,q}(g), \|\hat{\beta}_t^{\text{Bayes}}\|)}_{\text{conditional belief alignment}} \underbrace{\frac{\mathbb{E}[\hat{a}_{t,q}(g)^2 \mid \mathcal{C}_t]}{\mathbb{E}[\hat{a}_{t,q}(g)^2 + \varepsilon^2_{t,q} \mid \mathcal{C}_t]}}_{\text{conditional addressing fidelity}}. \tag{86}$$

The next two lemmas connect addressing capacity to regret: Lemma 21 bounds the average two-factor product by $\text{Cap}_t(g)/N_t$, and Lemma 22 shows that sublinear capacity forces linear regret. This reduces the regret floors in Theorem 3 to bounding each architecture's addressing capacity.

Lemma 21 (Capacity controls the two-factor product)**.** *Conditionally on $\mathcal{C}_t$, the product in (86), averaged over the upcoming key, is at most $\text{Cap}_t(g)/N_t$. Moreover,*

$$\frac{1}{N_t}\sum_{q\in\mathcal{U}_t} \mathbb{E}[(g_{t,q} - (\hat{\beta}_t^{\text{Bayes}})^\top X_q)^2 \mid \mathcal{C}_t] \geq \|\hat{\beta}_t^{\text{Bayes}}\|^2 \left(1 - \frac{\text{Cap}_t(g)}{N_t}\right). \tag{87}$$

Proof. When $\|\hat{\beta}_t^{\text{Bayes}}\| > 0$, multiplying Z_q by the $\mathcal{C}_t$-measurable nonzero scalar $\|\hat{\beta}_t^{\text{Bayes}}\|$ does not change conditional squared correlation. When $\|\hat{\beta}_t^{\text{Bayes}}\| = 0$, the two-factor product is zero. Thus each product is at most $\text{corr}^2_{\mathcal{C}_t}(g_{t,q}, Z_q)$, and averaging proves the first claim.

For the second claim, the conditional distance from the target $\|\hat{\beta}_t^{\text{Bayes}}\| Z_q$ to the line spanned by $g_{t,q}$ gives

$$\mathbb{E}[(g_{t,q} - \|\hat{\beta}_t^{\text{Bayes}}\| Z_q)^2 \mid \mathcal{C}_t] \geq \|\hat{\beta}_t^{\text{Bayes}}\|^2 \left(1 - \text{corr}^2_{\mathcal{C}_t}(g_{t,q}, Z_q)\right).$$

Use (84) and average over q. □

Lemma 22 (Sublinear capacity forces linear regret)**.** *Recall $T_\rho = T - R$. Suppose there is a deterministic set of $\Omega(T_\rho)$ prediction times at which $N_t \geq R/2$ and $\text{Cap}_t(g) \leq \varepsilon_R R$ almost surely for every g in a predictor class, where $\varepsilon_R \to 0$ is deterministic and uniform over these times and predictors. For fixed p and fixed $\widetilde{\Sigma}_\theta \succ 0$, the class has $\Omega(T_\rho)$ Bayes regret.*

Proof. Bayes orthogonality and the uniform reveal give

$$\ell_t(g) - \ell_t(g^{\text{Bayes}}) = \mathbb{E}\left[\frac{1}{N_t}\sum_{q\in\mathcal{U}_t} \mathbb{E}[(g_{t,q} - (\hat{\beta}_t^{\text{Bayes}})^\top X_q)^2 \mid \mathcal{C}_t]\right].$$

At the designated times, Lemma 21 yields

$$\ell_t(g) - \ell_t(g^{\mathrm{Bayes}}) \geq (1 - o(1))\mathbb{E}\|\hat{\beta}_t^{\mathrm{Bayes}}\|^2. \tag{88}$$

Recall the expected Bayes estimation error e_j^{Bayes} from (8). Total variance gives

$$\mathbb{E}\|\hat{\beta}_t^{\mathrm{Bayes}}\|^2 = \Lambda - e_{N_t^\rho}^{\mathrm{Bayes}}, \qquad \Lambda = \operatorname{tr}\widetilde{\Sigma}_\theta > 0. \tag{89}$$

The Bayes error estimate (21) yields

$$\sum_{j=0}^{T_\rho - 1} e_j^{\mathrm{Bayes}} = p\sigma^2 \log T_\rho + O(1). \tag{90}$$

The sum over any subset of matched times is no larger. Summing (88) over $\Omega(T_\rho)$ designated times therefore gives $\Lambda\Omega(T_\rho) - O(\log T_\rho) = \Omega(T_\rho)$. Notice that no factor pR appears: each key presents only one target direction, the projection of its candidate vector along the current Bayes belief. □

F.3.2 Bounds on addressing capacity

To apply Lemma 22, we develop capacity bounds for the architectural restrictions established in §F.3.3. Lemma 23 provides the common geometric reduction, used for linear spaces in Corollary 3 and polynomial images in Lemma 24. Lemma 25 supplies the conditional extension needed for selective SSMs.

Lemma 23 (Geometric representation of capacity)**.** *Let $G_t := (g_{t,q})_{q\in\mathcal{U}_t}$ be the prediction vector. For each $q \in \mathcal{U}_t$, define the optimal conditional least-squares rescaling of $g_{t,q}$ toward Z_q by*

$$\omega_{t,q}^\star := \frac{\mathbb{E}[Z_q g_{t,q} \mid \mathcal{C}_t]}{\mathbb{E}[g_{t,q}^2 \mid \mathcal{C}_t]}.$$

Set $\widetilde{G}_t := (\omega_{t,q}^\star g_{t,q})_{q\in\mathcal{U}_t}$, the rescaled prediction vector, and $Z := (Z_q)_{q\in\mathcal{U}_t}$. Then

$$\mathrm{Cap}_t(g) = N_t - \mathbb{E}[\|Z - \widetilde{G}_t\|^2 \mid \mathcal{C}_t]. \tag{91}$$

Consequently, if $\widetilde{G}_t$ lies almost surely in a $\mathcal{C}_t$-measurable set $\mathcal{I}_t \subseteq \mathbb{R}^{N_t}$ containing 0, then

$$\mathrm{Cap}_t(g) \leq N_t - \mathbb{E}[\mathrm{dist}^2(Z, \mathcal{I}_t) \mid \mathcal{C}_t]. \tag{92}$$

Proof. Conditional on $\mathcal{C}_t$, each Z_q has unit second moment and $\omega_{t,q}^\star g_{t,q}$ is its L^2 projection onto the line spanned by $g_{t,q}$. Hence

$$\mathbb{E}[(Z_q - \omega_{t,q}^\star g_{t,q})^2 \mid \mathcal{C}_t] = 1 - \mathrm{corr}^2_{\mathcal{C}_t}(g_{t,q}, Z_q).$$

Summing proves (91). The distance inequality follows pointwise from $\widetilde{G}_t \in \mathcal{I}_t$. □

Corollary 3 (Linear-subspace capacity bound)**.** *Suppose G_t lies almost surely in a $\mathcal{C}_t$-measurable linear subspace $V_t \subseteq \mathbb{R}^{N_t}$ of dimension at most K. Then $\mathrm{Cap}_t(g) \leq K$.*

Proof. Let $\Omega_t := \mathrm{diag}(\omega_{t,q}^\star : q \in \mathcal{U}_t)$. Then $\widetilde{G}_t = \Omega_t G_t$ lies in the subspace $\Omega_t V_t$, whose dimension is at most K. For a standard Gaussian Z and a fixed subspace V,

$$N_t - \mathbb{E}[\mathrm{dist}^2(Z, V) \mid \mathcal{C}_t] = \dim V.$$

Apply Lemma 23. □

Lemma 24 (Fixed-degree polynomial-image capacity bound)**.** *Fix an integer $k \geq 1$. Suppose G_t lies almost surely in the image of a $\mathcal{C}_t$-measurable polynomial map $Q_t : \mathbb{R}^m \to \mathbb{R}^{N_t}$ of degree at most k. Then*

$$\mathrm{Cap}_t(g) \leq C_k(m+1)\log(eN_t),$$

where C_k depends only on k.

Proof. Coordinatewise multiplication by Ω_t preserves both degree and parameter dimension, so $\widetilde{G}_t$ lies in the image of the polynomial map $\Omega_t Q_t$ of degree at most k. By the projection identity in Lemma 23, it suffices to prove that every polynomial map $Q : \mathbb{R}^m \to \mathbb{R}^N$ of degree at most k satisfies

$$N - \mathbb{E}\,\mathrm{dist}^2(Z, Q(\mathbb{R}^m)) \le C_k(m+1)\log(eN), \qquad Z \sim \mathcal{N}(0, I_N). \tag{93}$$

Write $Q_i(x) = \sum_{|\alpha|\le k} c_{i,\alpha} x^\alpha$ and define

$$\widetilde{Q}_i(s, x) := \sum_{|\alpha|\le k} c_{i,\alpha} s^{k-|\alpha|} x^\alpha.$$

Then $\widetilde{Q} : \mathbb{R}^{m+1} \to \mathbb{R}^N$ is homogeneous of degree k and $\widetilde{Q}(1, x) = Q(x)$. Let $\mathcal{D} := \overline{\widetilde{Q}(\mathbb{R}^{m+1})}$. This is a closed cone under nonnegative scaling because $c\widetilde{Q}(y) = \widetilde{Q}(c^{1/k} y)$ for $c \ge 0$, and it contains $Q(\mathbb{R}^m)$. If $\mathcal{D} = \{0\}$, the claim is immediate. Otherwise, with $\mathcal{S} := \mathcal{D} \cap \mathbb{S}^{N-1}$, optimization along each ray gives

$$N - \mathbb{E}\,\mathrm{dist}^2(Z, \mathcal{D}) = \mathbb{E} \sup_{u \in \mathcal{S}} \langle Z, u\rangle_+^2.$$

Let $\pi(x) := x/\|x\|$ for $x \ne 0$ and put

$$\mathcal{V} := \pi\Big(\widetilde{Q}(\mathbb{R}^{m+1}) \setminus \{0\}\Big).$$

If $u \in \mathcal{S}$, take y_j with $\widetilde{Q}(y_j) \to u$. Homogeneity gives

$$\widetilde{Q}\left(\frac{y_j}{\|\widetilde{Q}(y_j)\|^{1/k}}\right) = \frac{\widetilde{Q}(y_j)}{\|\widetilde{Q}(y_j)\|} \longrightarrow u.$$

Conversely $\mathcal{V} \subseteq \mathcal{S}$, so $\mathcal{S} = \overline{\mathcal{V}}$.

If $m + 1 > N$, then the desired estimate follows from the trivial bound

$$N - \mathbb{E}\,\mathrm{dist}^2(Z, Q(\mathbb{R}^m)) \le N \le (m+1)\log(eN).$$

Hence assume $m + 1 \le N$. By projectivized polynomial-image regularity (Zhang & Kileel, 2025, Lemma 2.15(c)), $\mathcal{V}$ is $((4k+1)^{m+2}, m+1)$-regular. Since $\mathcal{V} \subseteq \mathbb{S}^{N-1} \subseteq [-1, 1]^N$, the regular-set covering theorem (Zhang & Kileel, 2025, Theorem 2.6) gives

$$\log \mathcal{N}(\mathcal{V}, \varepsilon) \le C(m+1)\log\left(\frac{C(4k+1)N}{\varepsilon}\right)$$

for $0 < \varepsilon \le \mathrm{diam}(\mathcal{V})$; larger radii are trivial. Here and below C is universal. Passing to the closure changes the covering radius by at most a constant factor, and adjoining the negative set multiplies the covering number by at most two. Thus, for $\mathcal{S}_\pm := \mathcal{S} \cup (-\mathcal{S})$,

$$\log \mathcal{N}(\mathcal{S}_\pm, \varepsilon) \le C(m+1)\log\left(\frac{C(4k+1)N}{\varepsilon}\right).$$

Dudley's entropy integral therefore gives

$$\mathbb{E} \sup_{u \in \mathcal{S}_\pm} \langle Z, u\rangle \le C_k\sqrt{(m+1)\log(eN)}.$$

The supremum is a 1-Lipschitz function of Z, so Gaussian concentration upgrades this to the second-moment bound

$$\mathbb{E}\left[\left(\sup_{u \in \mathcal{S}_\pm} \langle Z, u\rangle\right)^2\right] \le C_k(m+1)\log(eN).$$

This bounds the preceding positive-part supremum. Enlarging $Q(\mathbb{R}^m)$ to $\mathcal{D}$ can only decrease distance, proving (93). □

The convolutional input to a selective transition can contain a linear summary of the candidate coordinates. The next lemma accounts for those directions before applying the polynomial-image bound to the remaining Gaussian coordinates.

Lemma 25 (Polynomial-image capacity bound with linear side information)**.** *Fix an integer $k \geq 1$. Conditionally on $\mathcal{C}_t$, let $S_t = W_t Z$, where the matrix W_t is $\mathcal{C}_t$-measurable and has rank at most r_{side}. Suppose that, conditional on $(\mathcal{C}_t, S_t)$, G_t lies in the image of a $(\mathcal{C}_t, S_t)$-measurable polynomial map $Q_{t,S_t} : \mathbb{R}^m \to \mathbb{R}^{N_t}$ of degree at most k. Then*

$$\mathrm{Cap}_t(g) \leq r_{\mathrm{side}} + C_k(m+1)\log(eN_t).$$

Proof. Let $r_t^{\mathrm{side}} := \mathrm{rank}(W_t)$. If $r_t^{\mathrm{side}} = N_t$ the claim is immediate. Otherwise, conditional on $\mathcal{C}_t$, apply a fixed Borel-measurable orthonormal-frame rule on each rank stratum to choose a $\mathcal{C}_t$-measurable isometry $V_t : \mathbb{R}^{N_t - r_t^{\mathrm{side}}} \to \ker W_t$. Then $Z_t^\circ := V_t^\top Z \sim \mathcal{N}(0, I_{N_t - r_t^{\mathrm{side}}})$ independently of S_t. The conditional moments defining Ω_t are $\mathcal{C}_t$-measurable, so for every realized $S_t = s$, the map $V_t^\top \Omega_t Q_{t,s}$ has degree at most k. Hence

$$\begin{aligned}\|Z - \Omega_t G_t\|^2 &\geq \|V_t^\top (Z - \Omega_t G_t)\|^2 \\ &\geq \mathrm{dist}^2\big(Z_t^\circ, V_t^\top \Omega_t Q_{t,S_t}(\mathbb{R}^m)\big).\end{aligned}$$

Condition on S_t, apply (93), and then average over S_t. This gives

$$\mathbb{E}[\|Z - \Omega_t G_t\|^2 \mid \mathcal{C}_t] \geq N_t - r_t^{\mathrm{side}} - C_k(m+1)\log(eN_t).$$

Use (91) and $r_t^{\mathrm{side}} \leq r_{\mathrm{side}}$. □

F.3.3 Architecture-specific geometry

We now establish the geometric restrictions needed to apply the bounds in §F.3.2 to our architecture classes (§2.1). Lemma 26 places the predictions of linear attention and fixed-transition or affine-gated SSMs in low-dimensional subspaces. For one-layer selective SSMs, the same lemma places the prediction vector in the image of a polynomial map after conditioning on a linear summary of the candidates, allowing us to apply Lemma 25. Lemma 27 extends the selective argument to our fixed-depth class. Our bounds also cover the count-normalized extensions of Appendix C.2, denoted by the superscript cal.

Lemma 26 (Architecture-specific geometry)**.** *Fix $t \geq R$ and condition on $\mathcal{C}_t$. The following statements hold for every predictor in the indicated classes, including their count-normalized extensions with any admissible schedule.*

(a) Linear attention and fixed-transition SSMs. *For a predictor in $\mathcal{F}_{\mathrm{att}}(D, d, H, L)$ or $\mathcal{F}_{\mathrm{SSM}}(D, d, H, d_{\mathrm{conv}})$, the prediction vector G_t lies in a $\mathcal{C}_t$-measurable linear subspace of dimension at most*

$$\binom{d_{\mathrm{key}} + 3^D}{3^D}.$$

(b) Affine-gated SSMs. *For a predictor in $\mathcal{F}^{\mathrm{aff}}_{\mathrm{SSM}}(D, d, H, d_{\mathrm{conv}})$, the prediction vector G_t lies in a $\mathcal{C}_t$-measurable linear subspace of dimension at most*

$$\binom{d_{\mathrm{key}} + 4^D}{4^D}.$$

(c) One-layer selective SSMs. *For a predictor in $\mathcal{F}^{\mathrm{sel}}_{\mathrm{SSM}}(1, d, H, d_{\mathrm{conv}})$, there is a linear summary of the candidate coordinates, $S_t = W_t Z$, with W_t measurable with respect to $\mathcal{C}_t$. Conditional on $(\mathcal{C}_t, S_t)$, the prediction vector G_t lies in the image of a $(\mathcal{C}_t, S_t)$-measurable polynomial map $Q_t : \mathbb{R}^m \to \mathbb{R}^{N_t}$ of degree at most four, with*

$$\mathrm{rank}\, W_t \leq Hd + 1, \qquad m \leq 6Hd.$$

The subspace dimension bounds in (a)–(b) depend only on D and d_{key}; the rank and dimension bounds in (c) are independent of d_e, d_{conv}, and p.

Proof. For attention, replace only the upcoming-key block at the current position by a variable $\chi \in \mathbb{R}^{d_{\mathrm{key}}}$. Causality makes every earlier position independent of χ. If the current representation has

degree δ_r in χ after layer r, a linear-attention cross term has degree at most δ_r, while the current-position self term — a bilinear score times a linear value — has degree at most $3\delta_r$. Residuals, head recombination, and the final linear readout do not increase degree. Since $\delta_0 = 1$, induction gives $\delta_D \le 3^D$. Evaluating the polynomial on the N_t public codewords places G_t in the fixed span of their degree-$\le 3^D$ monomial-evaluation vectors. Candidate vectors, including the state they induced earlier, affect only the coefficients of this polynomial and not that span.

For the fixed-transition and affine-gated SSM bounds, make the same substitution in the current token. Every earlier residual token and pre-update state is independent of χ by causality, although it may depend arbitrarily on Z. Suppose a layer input has coordinatewise degree at most m in χ. Its convolution has the same degree and its bilinear write has degree at most $2m$. With a fixed transition, the updated state has degree at most $2m$, so the ordinary read has degree at most $3m$. In the affine-gate ablation, the delta term $\beta_t k_t(k_t^\top h_{t-1})$ can have degree $3m$; uniformly over all four transition forms, the ordinary read therefore has degree at most $4m$. Head recombination and residual addition do not increase degree. Starting from degree one and evaluating on the codebook proves the SSM claims in parts (a) and (b).

The deterministic count-normalization factors are fixed scalars after conditioning on the prediction time. They do not change any of these degree bounds, for attention or for the SSM.

It remains to track only the candidate projections exposed to nonlinear gates. Conditional on $\mathcal{C}_t$, each head-specific first-layer convolved token has a decomposition

$$\widetilde{e}_{t,q}^{(1,b)} = \mu_{t,q}^{(b)} + B_t^{(b)} Z,$$

where $\mu_{t,q}^{(b)}$ and $B_t^{(b)}$ are $\mathcal{C}_t$-measurable. The rank of the raw map $B_t^{(b)}$ can grow with the ambient and convolution widths, so we do not condition on the complete convolved token. Instead, for each head retain only the candidate-dependent part of the affine preactivation to its sigmoid gate. This has dimension zero for a fixed transition, one for a scalar gate or delta strength, and at most d for a diagonal gate or generalized delta transition. Stacking the heads gives a linear summary of Z of rank at most Hd.

If $t+1$ begins a cycle, let q_t be the current key and append Z_{q_t} to this summary. Its candidate predecessor is the current regressor, so this additional coordinate fixes the candidate scalar carried directly by the current residual token. Away from a cycle boundary no candidate vector occurs in that token. Thus the total side-information rank is at most $Hd + 1$.

Conditional on this side information, every first-layer sigmoid value is a fixed coefficient for each counterfactual key. Treat as free variables the Hd pre-update state coordinates and the candidate-dependent parts of the three affine write/read projections, contributing $3Hd$ coordinates. A delta transition additionally requires the Hd candidate-dependent direction coordinates, while a generalized delta transition requires at most $2Hd$ coordinates for its two affine directions. Hence at most $6Hd$ free coordinates suffice uniformly over the permitted transition menu. The bilinear write has degree at most two. Fixed, scalar-gated, and diagonal-gated transitions preserve this degree before the ordinary read, while the delta and generalized delta terms have degree at most three in the free coordinates. The ordinary token–state read raises the uniform degree to at most four. Residual addition, head recombination, deterministic count normalization, and final prediction do not increase it. Collecting the coordinate polynomials over $q \in \mathcal{U}_t$ proves part (c). □

The one-layer argument conditions only on the candidate projections that feed sigmoid gates, rather than on the entire convolved token. When selectivity appears only in the first layer and every later transition is fixed, the same argument propagates through the entire stack:

Lemma 27 (Geometry of selective SSM stacks)**.** *Fix $D \ge 1$ and consider $\mathcal{F}_{\mathrm{SSM}}^{\mathrm{sel}}(D, d, H, d_{\mathrm{conv}})$, for arbitrary ambient width d_e and maximum convolution width d_{conv}, including its count-normalized extensions with any admissible schedule. The first layer uses an arbitrary permitted convolution width and any permitted transition, while every later layer uses an arbitrary permitted convolution width and a fixed transition. Its writes are bilinear and its reads are ordinary token–state reads. Its D layers contain DHd recurrent coordinates. There is a linear summary of the candidate coordinates, $S_t = W_t Z$, with W_t measurable with respect to $\mathcal{C}_t$ and*

$$\operatorname{rank} W_t \le Hd + 1$$

such that, conditionally on $(\mathcal{C}_t, S_t)$, G_t lies in the image of a degree-$\le 4 \cdot 3^{D-1}$ polynomial map $Q_t : \mathbb{R}^m \to \mathbb{R}^{N_t}$ with $m \le (4D+2)Hd$.

Proof. Use the projection-specific side information constructed in Lemma 26(c) for the first layer. Conditional on $(\mathcal{C}_t, S_t)$, all of its sigmoid values are fixed coefficients for each counterfactual key, and the direct current-token residual is fixed even at a cycle boundary.

Retain as free variables all DHd pre-update state coordinates. From the candidate-dependent part of the first-layer convolved tokens, retain the three Hd-dimensional projections entering the bilinear write and ordinary read. For delta transitions also retain the Hd projected direction coordinates; for generalized delta transitions retain the at most $2Hd$ coordinates entering their two affine directions. The scalar and diagonal transitions require no direction variables. Thus the first layer contributes at most $5Hd$ projected coordinates beyond its pre-update state.

For each later-layer head (r, b), let

$$\eta_t^{(r,b)} := \sum_{j=1}^{d_{\mathrm{conv}}-1} C_j^{(r,b)} \odot e_{t-j}^{(r-1)}, \qquad r = 2, \dots, D,$$

where an empty sum is zero. This is the entire lagged part of the convolution and is shared across the counterfactual upcoming keys. Retain as free parameters

$$\bar{W}_K^{(r,b)} \eta_t^{(r,b)}, \qquad \bar{W}_V^{(r,b)} \eta_t^{(r,b)}, \qquad W_Q^{(r,b)} \eta_t^{(r,b)}, \qquad r = 2, \dots, D,$$

where bars denote the linear parts of the affine write maps. Together with the preceding first-layer projections and all pre-update states, these form a vector ϑ_t of dimension at most

$$DHd + 5Hd + 3(D-1)Hd = (4D+2)Hd.$$

Its realized value may depend arbitrarily and nonlinearly on all candidate vectors; treating it as free only enlarges the set of possible prediction vectors.

The first-layer bilinear write has degree at most two in ϑ_t. A fixed, scalar-gated, or diagonal-gated transition leaves the updated state at degree at most two, whereas a delta or generalized delta term has degree at most three. Multiplication by the affine ordinary-read selector therefore gives $e_{t,q}^{(1)}$ degree at most four, uniformly over the permitted transition menu. Put $\delta_1 := 4$ and suppose the current representation after layer $r-1$ has degree at most δ_{r-1}, where $r \ge 2$. The current-token part of each layer-r convolution is linear in $e_{t,q}^{(r-1)}$, while its lagged part enters the write factors and read selector only through the three retained projections above. Each such factor is therefore a polynomial of degree at most δ_{r-1}. The bilinear write has degree at most $2\delta_{r-1}$, as does the updated state under the fixed transition. The ordinary token–state read consequently has degree at most $3\delta_{r-1}$, and residual addition and head recombination do not increase it. Hence $\delta_r \le 3\delta_{r-1} \le 4 \cdot 3^{r-1}$. Deterministic count normalization and the final prediction map also preserve degree. Collecting the coordinate polynomials over $q \in \mathcal{U}_t$ proves the claim. Later convolution widths affect the realized lag aggregates, but neither their three projected dimensions per head nor the degree bound. □

Scope of the selective-SSM bound. Our class $\mathcal{F}_{\mathrm{SSM}}^{\mathrm{sel}}$ permits selectivity only in the first layer. This lets us fix sigmoid gate outputs by conditioning on linear candidate projections, preserving Gaussianity in the remaining directions. The inputs to deeper sigmoid gates generally depend nonlinearly on the candidates: conditioning on them need not preserve Gaussianity, while leaving the gates unconditioned generally breaks the polynomial structure used above. Extending the bound to fully selective stacks is therefore beyond the present argument.

F.3.4 Proof of Theorem 3

We now combine the preceding bounds to prove Theorem 3. Lemma 26 and Corollary 3 yield the attention and affine-gate bounds in (a)–(b), while Lemmas 25 and 27 yield the selective-SSM bound in (c). Lemma 22 then gives the regret floors. We use the main-text class notation below; all bounds also hold uniformly for the count-normalized extensions of Appendix C.2, under every admissible schedule.

Lemma 28 (Fixed-depth addressing floors)**.** *Let $d_{\mathrm{key}} = \Theta(\log R)$ and fix D independently of R. For every choice of d_e, d, H, L, d_{conv}, and $t \geq R$, every predictor g in $\mathcal{F}_{\mathrm{att}}(D, d, H, L)$ or $\mathcal{F}_{\mathrm{SSM}}(D, d, H, d_{\mathrm{conv}})$ obeys*

$$\mathrm{Cap}_t(g) \leq \min\left\{N_t, \binom{d_{\mathrm{key}} + 3^D}{3^D}\right\}. \tag{94}$$

Every predictor $g \in \mathcal{F}^{\mathrm{aff}}_{\mathrm{SSM}}(D, d, H, d_{\mathrm{conv}})$ with the same dimensions obeys

$$\mathrm{Cap}_t(g) \leq \min\left\{N_t, \binom{d_{\mathrm{key}} + 4^D}{4^D}\right\}. \tag{95}$$

Consequently, in the main-text regime $R \to \infty$, $T_\rho \to \infty$, and $R = o(\sqrt{T_\rho})$, every such fixed-depth predictor has $\Omega(T_\rho)$ regret, uniformly over d_e, d, H, L, d_{conv}, and parameter choices.

Proof. Lemma 26 and Corollary 3 give the displayed bounds. Each degree is constant at fixed D, so every binomial term is $\mathrm{polylog}(R) = o(R)$. The first half of each complete post-initial cycle has $N_t \geq R/2$; these account for $T_\rho/2 - O(R) = \Omega(T_\rho)$ prediction times. Lemma 22 gives the regret floor. □

Before proving the fixed-depth state requirement in Lemma 29, we illustrate the argument for one-layer selective SSMs. Lemma 26(c) and Lemma 25 with $k = 4$ give

$$\mathrm{Cap}_t(g) \leq Hd + 1 + C_4(6Hd + 1)\log(eN_t) \leq C(Hd + 1)\log(eN_t). \tag{96}$$

Here C is universal, and the additive one is needed only at cycle boundaries. The bound holds uniformly over the suppressed ambient width d_e. The next lemma extends this argument to our fixed-depth selective SSM class, expressing the bound in terms of the total recurrent-state dimension DHd.

Lemma 29 (Fixed-depth selective SSM state floor)**.** *Fix D independently of R and consider the depth-D class in Lemma 27. Uniformly over the public codebook, ambient width d_e, maximum convolution width d_{conv}, all other resources, and all parameter choices,*

$$\mathrm{Cap}_t(g) \leq C_D\, DHd \log(eN_t), \qquad t \geq R, \tag{97}$$

where C_D depends only on D. In the regime $R \to \infty$, $T_\rho \to \infty$, and $R = o(\sqrt{T_\rho})$,

$$DHd = o(R/\log R) \quad \Longrightarrow \quad \mathrm{Reg}_T(g) = \Omega(T_\rho). \tag{98}$$

Consequently, every predictor sequence in this fixed-depth class with $o(T_\rho)$ regret satisfies $DHd = \Omega(R/\log R)$.

Proof. Lemma 27 gives degree at most $4 \cdot 3^{D-1}$, parameter dimension at most $(4D + 2)Hd$, and linear side-information rank at most $Hd + 1$. Lemma 25 with $k = 4 \cdot 3^{D-1}$ therefore gives

$$\mathrm{Cap}_t(g) \leq Hd + 1 + C_{4\cdot 3^{D-1}}\big((4D + 2)Hd + 1\big)\log(eN_t) \leq C_D\, DHd \log(eN_t),$$

which proves (97).

Now restrict to times for which $t + 1$ lies in positions $2, \ldots, \lfloor R/2 \rfloor$ of a complete post-initial cycle. There are $\Omega(T_\rho)$ such times, none is a cycle boundary, and $N_t \geq R/2$ on all of them. If $DHd = o(R/\log R)$, then the capacity bound is $o(R)$ uniformly on these times. Lemma 22 proves (98).

Finally, if an $o(T_\rho)$-regret sequence failed the stated $\Omega(R/\log R)$ necessity, some subsequence would satisfy $DHd = o(R/\log R)$, contradicting (98). □

Proof of Theorem 3. At fixed depth, the capacity bounds in Lemma 28 are $o(R)$, proving parts (a)–(b). Since $N_t \leq R$, Lemma 29 gives part (c). Their regret conclusions complete the proof. □

F.3.5 Constructions for Proposition 4

This section proves Proposition 4. Each construction first retrieves the required regressors, then applies its Case 1 kernel for belief maintenance. Proposition 11 gives their leading regret constants and sufficient dimensions.

Recall the Case 1 notation from (4)–(17):

$$v_i = \lambda_i + \operatorname{tr}\widetilde{\Sigma}_\theta + \sigma^2, \qquad \Lambda = \sum_i \lambda_i, \qquad V = \sum_i v_i,$$
$$\zeta = \sum_i \sqrt{\lambda_i v_i}, \qquad \bar{\zeta} = \sqrt{\Lambda V}.$$

We round $L^\star(T_\rho)$ to the nearest integer when choosing the window.

Proposition 11 (Content-routing pair)**.** *Assume $R = o(\sqrt{T_\rho})$ and keep p, $\widetilde{\Sigma}_\theta \succ 0$, and $\sigma^2 > 0$ fixed. The following depth-two predictors can be realized with ambient width $d_e = 2d_{\mathrm{key}} + 4p + 2$.*

(a) Attention. *A predictor $g_{\mathrm{att}} \in \mathcal{F}^{\mathrm{sm}}_{\mathrm{att}}(2, d_{\mathrm{key}} \vee p, 2, L)$ with $L = L^\star(T_\rho) + O(1)$ uses softmax in its first layer and linear attention in its second, and satisfies*

$$\mathrm{Reg}_T(g_{\mathrm{att}}) = \frac{2}{\sqrt{3}}\bar{\zeta}\sqrt{T_\rho} + o(\sqrt{T_\rho}).$$

(b) Selective SSM. *A predictor $g_{\mathrm{SSM}} \in \mathcal{F}^{\mathrm{sel}}_{\mathrm{SSM}}(2, 2R, p, 2)$ uses $4pR$ recurrent-state coordinates and satisfies*

$$\mathrm{Reg}_T(g_{\mathrm{SSM}}) = \zeta\sqrt{T_\rho} + o(\sqrt{T_\rho}).$$

We establish the attention and selective-SSM claims in turn.

Attention construction. The first layer uses two softmax heads to retrieve the regressors for evidence assembly and addressing. The second layer uses one active linear-attention head to apply the Case 1 uniform kernel; its other head is zero. We bound the error from finite softmax scores before deriving the regret.

Proof of Proposition 11(a). At position i, the first layer's atom head uses

$$s^{\mathrm{a}}_{ij} := \beta_{\mathrm{sm}}\langle\psi_i, \psi_j\rangle - \gamma(i-j) + p^{\mathrm{a}}_{i-j}, \qquad p^{\mathrm{a}}_0 = -M_0, \quad p^{\mathrm{a}}_u = 0\ (u > 0), \tag{99}$$

and copies ξ_j. The lag penalty $-\gamma(i-j)$ favors more recent occurrences of the same key. The finite lag-zero penalty prevents the current token, which has the same key and zero recency cost, from winning. The query head uses

$$s^{\mathrm{q}}_{ij} := \beta_{\mathrm{sm}}\langle\psi_{i+1}, \psi_j\rangle - \gamma(i-j) \tag{100}$$

and also copies ξ_j. Denote the routed residual coordinates by $\widetilde{u}_i$ and $\widetilde{q}_i$, respectively.

Let $\eta_0 := \max_{r\neq s} |\langle\psi^{(r)}, \psi^{(s)}\rangle| < 1/3$. For any fixed $a > 0$, choose finite parameters satisfying

$$\gamma \geq (1+a)\log T, \qquad \beta_{\mathrm{sm}}(1-\eta_0) - 2\gamma R \geq (1+a)\log(LT), \qquad M_0 - 2\gamma R \geq (1+a)\log(LT). \tag{101}$$

The desired previous occurrence is at lag at most $2R-1$, so it lies in the window whenever $L \geq 2R$. If δ_i is the mass outside the desired source in either head, the same-key recency gap and wrong-key content gap give uniformly

$$\delta_i \leq \bar{\delta}_T := \frac{e^{-\gamma}}{1-e^{-\gamma}} + L\exp[-\{\beta_{\mathrm{sm}}(1-\eta_0) - 2\gamma R\}] + \exp[-(M_0 - 2\gamma R)]. \tag{102}$$

The parameters in (101) may be increased so that $T^m\bar{\delta}_T \to 0$ for any prescribed fixed m.

The second layer uses linear attention with zero recency slope and its own parameters. Its active head uses score and value

$$s^{(2)}_{ts} := \frac{1}{L}\widetilde{q}_t^\top\widetilde{u}_s, \qquad W^{(2)}_V e^{(1)}_s = y_s e_{\mathrm{acc}}. \tag{103}$$

Here $e_{\mathrm{acc}} \in \mathbb{R}^d$ is a standard basis vector in the head's value space. The output projection $W_O^{(2)}$ maps it to the residual accumulator coordinate, which is zero after the first layer and selected by W_{pred}. Under exact routing, its scalar output is

$$g_t^{\mathrm{exact}} = \xi_{\rho(t+1)}^\top \left(\frac{1}{L} \sum_{s \in I_t} \xi_{\rho(s)} y_s \right). \tag{104}$$

Here $I_t := J_t \cap \{R+1, \dots, t\}$ is the matched portion of the physical length-L attention window, truncated during warm-up. The ambient width $d_e = 2d_{\mathrm{key}} + 4p + 2$ supplies two key blocks, the current and next transformed regressors, two routed vector blocks, and label/accumulator coordinates. The first-layer matching scores require d_{key} query–key coordinates, while copying a regressor and computing the second-layer inner product each require p value or query–key coordinates. Because these maps are independent, the sufficient common per-head width is $d_{\mathrm{key}} \vee p$.

It remains to control soft routing. Conditional on the key stream, a retrieved vector has the form

$$\widetilde{X} = (1-\delta) X_* + e_{\mathrm{off}}, \qquad e_{\mathrm{off}} = \sum_{j \neq *} \alpha_j X_j.$$

The off-target regressors are centered and independent of X_*, so

$$\mathbb{E}[e_{\mathrm{off}} X_*^\top \mid \text{keys}] = 0, \qquad \mathbb{E}[\|e_{\mathrm{off}}\|^4 \mid \text{keys}] \le C_p \delta^4.$$

Likewise, for a routed atom, $y\widetilde{X} = (1-\delta) y X_* + y e_{\mathrm{off}}$ with $\mathbb{E}[y e_{\mathrm{off}} \mid \beta, \text{keys}] = 0$. Fixed $p, \widetilde{\Sigma}_\theta, \sigma^2$ give uniform fourth moments. The uniform-kernel weights in (103) have total absolute mass at most one, so Minkowski's and Hölder's inequalities bound the increase in cumulative squared loss by $O(T\bar{\delta}_T) = o(\sqrt{T_\rho})$.

For $t \le R$, the attention output and the Bayes predictor are both zero. At $t = R + j$, $1 \le j \le T_\rho$, the window I_t contains $\min(j, L)$ matched atoms. Conditional on the keys, their routed regressors and the next query have exactly the Case 1 joint law. Thus the exact-routed regret is the Case 1 uniform-kernel regret over T_ρ steps. The kernel error identity gives the cost before subtracting the Bayes baseline:

$$\frac{\Lambda L}{3} + \frac{V T_\rho}{L} + O(1). \tag{105}$$

The posterior subtraction is $O(\log T_\rho) = o(\sqrt{T_\rho})$. Taking L as the nearest integer to $L^\star(T_\rho)$ therefore gives $\frac{2}{\sqrt{3}} \bar{\zeta} \sqrt{T_\rho} + o(\sqrt{T_\rho})$. Since $R = o(\sqrt{T_\rho})$ and $L^\star(T_\rho) = \Theta(\sqrt{T_\rho})$, the reach condition $L \ge 2R$ is eventually satisfied. Together with the soft-routing error bound, this proves Proposition 11(a). □

Selective-SSM construction. The first layer caches regressors in two state blocks per eigencoordinate and reads the regressor named by the upcoming key into the residual stream. The second layer uses a width-two convolution to pair the preceding cache read with the current label, then applies the Case 1 exponential kernel. We verify the cache before implementing the layers and bounding regret.

The first layer uses $H = p$ heads. Head i handles the scalar eigencoordinate $\xi_{i,t} := e_i^\top \xi_t$ and stores

$$\bar{h}_{i,t} = (C_{i,t}^{(1)}, C_{i,t}^{(2)}) \in \mathbb{R}^{2R}, \tag{106}$$

where the two state blocks each have one coordinate per key. The first block keeps the latest regressor for each key relative to a common reference. The second uses the same writes but also resets the upcoming key, so their difference isolates the regressor needed for prediction. Let $\iota_t \in [R]$ index the key $\psi_t = \psi^{(\iota_t)}$. The codewords enter the affine gate maps as parameters, not recurrent-state coordinates.

In the hard-gate limit, the two state blocks use the following gates, where 1 retains the previous coordinate value and 0 clears it before the new write:

$$\lambda_{t,r}^{(1)} = \mathbf{1}\{r \neq \iota_t\}, \qquad \lambda_{t,r}^{(2)} = \mathbf{1}\{r \notin \{\iota_t, \iota_{t+1}\}\}. \tag{107}$$

Both are limits of sigmoid gates in our class. At steepness s, take

$$\lambda_{t,r}^{(1,s)} := \sigma\left(s \left\{ \frac{2}{3} - \langle \psi^{(r)}, \psi^{(\iota_t)} \rangle \right\} \right), \qquad \lambda_{t,r}^{(2,s)} := \sigma\left(s \left\{ \frac{2}{3} - \langle \psi^{(r)}, \psi^{(\iota_t)} + \psi^{(\iota_{t+1})} \rangle \right\} \right). \tag{108}$$

Each gate is the sigmoid of an affine function of the token's two key blocks. The coherence bound gives a margin of at least $\kappa_R := 1/3 - \eta_0 > 0$ on either side of $2/3$, including when the two keys coincide. The bilinear write, with one constant affine factor, broadcasts $u_{i,t} := \xi_{i,t} - \xi_{i,t+1}$ to both state blocks:

$$C^{(1)}_{i,t,r} = \lambda^{(1)}_{t,r} C^{(1)}_{i,t-1,r} + u_{i,t}, \tag{109}$$

$$C^{(2)}_{i,t,r} = \lambda^{(2)}_{t,r} C^{(2)}_{i,t-1,r} + u_{i,t}, \tag{110}$$

With hard gates, if key r last occurred at $\ell \le t$, telescoping gives $C^{(1)}_{i,t,r} = \xi_{i,\ell} - \xi_{i,t+1}$. The two state blocks agree outside the upcoming key: both reset the current key, erasing their previous difference. If $\iota_t \neq \iota_{t+1}$, the upcoming slot is $\xi_{i,\rho(t+1)} - \xi_{i,t+1}$ in the first block and $\xi_{i,t} - \xi_{i,t+1}$ in the second. If the keys coincide, both blocks reset the same slot and $\rho(t+1) = t$. Thus, with $\bar{c}^{\mathrm{cache}}_i := (\mathbf{1}_R, -\mathbf{1}_R)$, the residual addition of $\xi_{i,t}$ gives

$$\chi_{i,t} := \xi_{i,t} + \bar{c}^{\mathrm{cache}\top}_i \bar{h}_{i,t} = \xi_{i,\rho(t+1)} \qquad (t \ge R). \tag{111}$$

This includes equal consecutive keys across cycle boundaries.

For finite sigmoids, steepness $C\kappa_R^{-1}\log(RT)$ makes the uniform gate error δ_T an arbitrarily small fixed inverse power of RT. Let superscripts "soft" and "hard" denote the corresponding trajectories. With hard gates, entries in each state block are differences of two Gaussian coordinates, including before a key's first occurrence. Induction in (109)–(110), using gate values in $[0,1]$, therefore gives uniformly for $t \le T$

$$\max_r \left\| C^{(1),\mathrm{soft}}_{i,t,r} - C^{(1),\mathrm{hard}}_{i,t,r} \right\|_8 + \max_r \left\| C^{(2),\mathrm{soft}}_{i,t,r} - C^{(2),\mathrm{hard}}_{i,t,r} \right\|_8 \le C_{p,\widetilde{\Sigma}_\theta,\sigma} T \delta_T.$$

Summing these errors over the R keys bounds each routed-coordinate error by $C_{p,\widetilde{\Sigma}_\theta,\sigma} RT\delta_T$ in L^8.

Proof of Proposition 11(b). Work in the whitened eigencoordinates of (81) and use the linear embedding

$$e^{(0)}_t = (\psi_{t+1}, \psi_t, \xi_{t+1}, \xi_t, y_t, \xi_t, \xi_t, 0) \in \mathbb{R}^{2d_{\mathrm{key}}+4p+2}.$$

Layer 1 uses identity convolution and the coordinatewise sigmoid transitions of the cache recurrence above. A constant read selector and linear output projection sum the coordinatewise differences between each head's two state blocks and add the result to the corresponding coordinate in the two extra ξ_t blocks, leaving all other blocks unchanged. These residual blocks therefore contain two copies of $\chi_t = (\chi_{1,t}, \dots, \chi_{p,t})$. The copies occupy residual-stream coordinates and require no additional recurrent state.

The width-two convolution in layer 2 exposes y_t, χ_t, and χ_{t-1} in separate coordinates. Head i uses the bilinear write $y_t \chi_{i,t-1}$ and the fixed scalar recurrence

$$a_{i,t} = \alpha_i a_{i,t-1} + y_t \chi_{i,t-1}.$$

Each layer-2 head has $2R$ state coordinates, of which only one is active; the others remain zero. For $t \ge R+1$, the cache identity gives $\chi_{i,t-1} = \xi_{i,\rho(t)}$. Reading the updated state with $\chi_{i,t}$ and recombining the heads with coefficients $1-\alpha_i$ therefore yields

$$g_t = \sum_{i=1}^{p} (1-\alpha_i) a_{i,t} \chi_{i,t}. \tag{112}$$

The second-layer output projection writes only to the initially zero prediction coordinate, which W_{pred} selects. In the hard-gate limit this is the Case 1 exponential predictor in each direction over the matched steps. The first cycle's zero labels keep the accumulator and prediction zero, even with finite gates. At $t = R$, the Bayes belief is still zero; at the first matched write $t = R+1$, the identity for $\chi_{i,R}$ applies. The update-then-read convention incorporates y_t before predicting y_{t+1}.

Choose the decays by

$$W_i := \frac{1+\alpha_i}{1-\alpha_i} = 2\sqrt{\frac{v_i T_\rho}{\lambda_i}}. \tag{113}$$

The Case 1 error identity then gives, in direction i,

$$\frac{\lambda_i W_i}{4} + \frac{v_i T_\rho}{W_i} + o(\sqrt{T_\rho}) = \sqrt{\lambda_i v_i T_\rho} + o(\sqrt{T_\rho}).$$

Summing the fixed number of directions and subtracting the $O(\log T_\rho)$ posterior baseline gives $\zeta\sqrt{T_\rho} + o(\sqrt{T_\rho})$ for the hard-gate construction.

For finite gates, the layer-1 cache follows the soft trajectory above. The routed-coordinate error is at most $C_{p,\widetilde{\Sigma}_\theta,\sigma} RT\delta_T$ in L^8, and the labels have uniformly bounded L^8 norm, so Hölder's inequality gives the same order in L^4 for the layer-2 evidence-write error. Iterating the accumulator for at most T steps bounds its soft–hard error by $C_{p,\widetilde{\Sigma}_\theta,\sigma} RT^2\delta_T$ in L^4. The hard accumulator has L^4 norm at most $C_{p,\widetilde{\Sigma}_\theta,\sigma} T$ under a crude finite-horizon bound. A final Hölder bound in (112), followed by summation over time, bounds the increase in cumulative loss by $C_{p,\widetilde{\Sigma}_\theta,\sigma} RT^5\delta_T$. Taking $\delta_T \le (RT)^{-8}$ makes this $o(\sqrt{T_\rho})$, and the strict coherence margin attains this error with finite sigmoid slopes.

Both layers use bilinear writes and ordinary token–state reads, without count normalization ($\mathbf{c} = (1, 1)$ in the extended class). Finally, both layers have p heads of width $2R$, totaling $4pR$ recurrent coordinates. Their convolution widths $(1, 2)$ require d_e buffer coordinates, and the embedding above gives the stated ambient width. □

Specializing Lemma 29 to $D = 2$ pairs this construction with a nearly matching lower bound. Within the two-layer selective SSM class, sublinear regret requires $\Omega(R/\log R)$ recurrent coordinates, while the construction above uses $4pR = O(R)$ recurrent coordinates for fixed p and attains $\zeta\sqrt{T_\rho} + o(\sqrt{T_\rho})$ regret. It is therefore optimal in total recurrent state up to a logarithmic factor.