LOCALITY AND RETRIEVAL SPEND ONE BUDGET: A CONSERVATION LAW FOR SEQUENCE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Every causal, time-invariant sequence mixer obeys an identity its designers did not put there. With H the transfer function, $h[m]$ the first nonzero tap and $\mathcal{B}[H]$ the excess phase of the non-minimum-phase zeros, the frequency-averaged log-gain satisfies $\mathcal{L}[H] = \log|h[m]| + \mathcal{B}[H]$. It holds to 10^{-15} across 162 kernels drawn from S4, S4D, S5, LRU, RetNet, gated linear attention with a frozen gate, and plain causal convolutions, and to 10^{-18} on the released weights of Mamba-130M, 1.4B and 2.8B.

The identity is a conservation law. Local sharpness and long-range excess phase draw on one budget, and a retrieval head spends the whole of it on phase: we measure $\log|h[m]| = -2.82$ against $\mathcal{B} = +2.82$, summing to the all-pass ceiling $\mathcal{L} = 0$. The law forces a frontier. Nulling a fraction s of the band to depth ε costs $\varepsilon \geq |h[0]|^{1/s}(1-s)^{1/2s}$, which 45,000 random kernels never cross. Because the frontier follows from *time invariance* and not from state size, an input-dependent gate does not find a better point inside it; it leaves the space where the frontier lives.

That departure buys $+0.50$ accuracy on multi-query associative recall, at the intermediate state sizes where the frontier binds and nowhere else, across 48 matched runs. **A time-invariant model initialised over a spread of timescales recovers most of the gain**, which locates the operative ingredient in multi-timescale capacity rather than in gating, and makes it available with no extra parameters and no cost at inference. Beyond the boundary the selective class carries an invariant of its own, a contraction sum rule we verify to 5×10^{-8} on delta-rule transitions.

1 INTRODUCTION

The design space of subquadratic sequence mixers has been mapped by search. State size grew (S4 $\rightarrow$ S5 $\rightarrow$ Mamba-2), initialisations were engineered (HiPPO), gates were made input-dependent (Mamba, GLA), and each step was justified downstream. Missing from all of it is a statement of what the space *is*: a constraint obeyed by every member, unmoved by training, against which a design decision reads either as a move inside the space or as a departure from it.

Control theory has had such a statement since 1945. Bode (1945) showed that the log-magnitude of a causal, stable feedback loop, integrated over frequency, is pinned by the loop's poles, so that attenuating disturbances in one band forces amplification in another. Engineers call it the waterbed effect (Stein, 2003; Emami-Naeini & de Roover, 2019); in the form we need it is older still, being Jensen's formula. Nobody appears to have pointed it at a sequence model. A survey of 17,954 papers accepted at ICLR 2026, NeurIPS 2025 and ICML 2026 (Appendix B) turns up no dispersion relation and no frequency-domain sum rule anywhere in the machine-learning literature.

Pointing it at one turns out to explain something the sequence-modelling literature has circled empirically for three years.

Contributions. Figure 1 states the mechanism; the rest of the paper makes each part of it exact.

- **An exact invariant** (Thm. 1) of every causal LTI mixer, verified numerically to machine precision across S4D/S5/LRU, HiPPO-LegS, FIR/Hyena, RetNet decay and hand-built retrieval heads (Fig. 2).

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

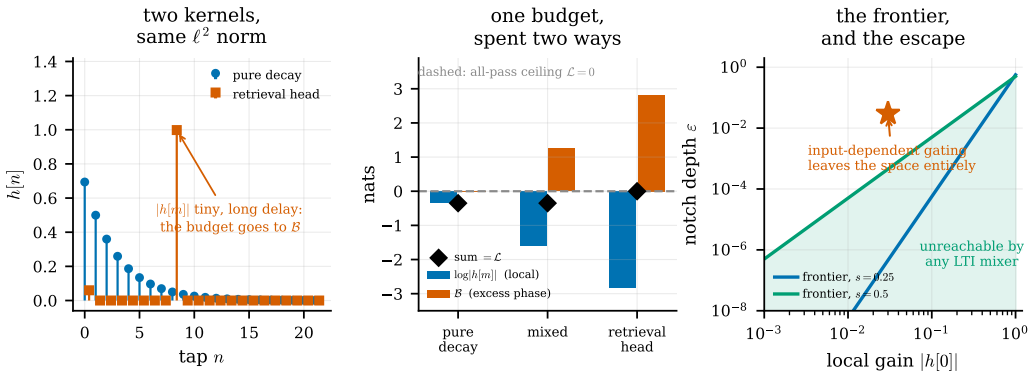


Figure 1: **The conservation law, and what it costs. Left:** two causal kernels of identical ℓ^2 norm. The decay head puts its mass in the first tap; the retrieval head has a tiny first tap and a long delay, and that delay is excess phase $\mathcal{B}$. **Middle:** Theorem 1 splits the frequency-averaged log-gain into a local term $\log|h[m]|$ and a phase term $\mathcal{B}$, so the two are alternative ways of spending one budget. A retrieval head is pushed to the all-pass ceiling $\mathcal{L}=0$, and that is where we measure it (-2.82 against $+2.82$). **Right:** the resulting frontier (Theorem 2): local gain and stop-band depth cannot both be small. The frontier follows from time invariance alone, so an input-dependent gate does not sit at a better point inside the region. It sits outside it.

Table 1: Architectures covered, exactly or in their frozen-gate limit.

family	kernel $h[n]$	instances
diagonal SSM	$\sum_{j\le N} c_j\lambda_j^n,\ \lvert\lambda_j\rvert<1$	S4D, S5, LRU, Mamba-2 (const. Δ)
scalar decay	γ^n	RetNet, GLA (frozen gate)
FIR	finite support	causal conv, Hyena (frozen filter)
orthogonal-polynomial	HiPPO-LegS discretised	S4

- **A frontier** (Thm. 2) trading local gain against notch depth, with a matching construction, empirically unviolated in $45{,}000$ kernels.
- **A checkpoint-computable measurement** $\mathrm{SEL}(h)$, exactly zero for every time-invariant mixer, quantifying how far a model's frozen kernels disperse in excess phase, with a sharp characterisation of its range (Rem. 5).
- **A decomposition of what selectivity buys** (§5): over 48 matched runs, input-dependent gating gains $+0.50$ accuracy where the frontier binds and nowhere else, and a time-invariant model with a spread of timescales recovers most of it, which isolates multi-timescale capacity as the operative ingredient.
- **An exact invariant for the selective class** (Prop. 6): a contraction sum rule computable from one forward pass, verified to 5×10^{-8} on delta-rule transitions.

Why an identity and not a bound. A bound invites the reader to ask how tight it is. An identity does not: $\mathcal{L}-\log|h[m]|-\mathcal{B}$ is zero, and a measurement of it is a test of whether the object being analysed is the object the analyst thinks it is. Our first experiment is that test, and it is why the rest can be trusted.

2 Setup

A *causal sequence mixer* maps $u\mapsto y$ by $y[n]=\sum_{k\ge0}h[k]u[n-k]$ with $h\in\ell^2(\mathbb{Z}_{\ge0})$. Table 1 lists the members of this class. Nonlinear mixers enter through their first-order Volterra kernel, i.e. the linearisation at an operating point; softmax attention is *not* in the class, and we return to this in §5.

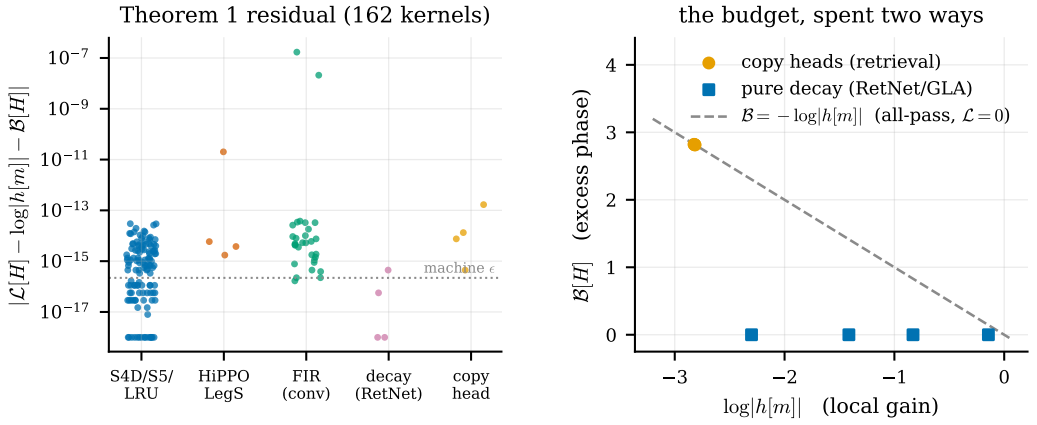


Figure 2: **Left:** residual of Eq. (2) for 162 kernels across five families; median 1.1×10^{-15}, i.e. machine precision. The single outlier at 1.7×10^{-7} is an FIR kernel with a zero within 10^{-4} of the unit circle, where the quadrature of the integrable log singularity is the limiting error, not the theorem. **Right:** the two ways of spending the budget. Pure-decay kernels (RetNet/GLA) are minimum phase: $\mathcal{B} = 0$, all budget in $h[0]$. Retrieval heads sit on the line $\mathcal{B} = -\log|h[m]|$, i.e. $\mathcal{L} = 0$: they have spent *everything* on excess phase and sit at the all-pass ceiling.

Write $H(z) = \sum_{n \geq 0} h[n] z^{-n}$ and $\hat{H}(\omega) = H(e^{i\omega})$; stability means the spectral radius $\rho(h) < 1$. Define

$$\mathcal{L}[H] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log|\hat{H}(\omega)|\, d\omega, \qquad \mathcal{B}[H] = \sum_{k:\,|w_k|<1} \log \frac{1}{|w_k|} \geq 0, \tag{1}$$

where w_k are the zeros, inside the unit disc, of $\tilde{H}(w) = \sum_n h[n] w^n$, equivalently the zeros of H *outside* the unit circle, the non-minimum-phase zeros. $\mathcal{B}$ is the log of the Blaschke factor; in control language it is the mixer's *excess phase*, the delay it carries beyond the minimum compatible with its magnitude response. Let $m = \min\{n : h[n] \neq 0\}$ be the intrinsic delay.

3 An Exact Invariant

Theorem 1 (Sum rule)**.** *Let h be causal with $\rho(h) < 1$, $h \not\equiv 0$. Then*

$$\boxed{\mathcal{L}[H] \;=\; \log|h[m]| \;+\; \mathcal{B}[H]} \tag{2}$$

and in particular $\mathcal{L}[H] \geq \log|h[m]|$, with equality iff H is minimum phase.

Proof. Set $w = 1/z$, $\tilde{H}(w) = \sum_{n \geq m} h[n] w^n = w^m G(w)$ with $G(0) = h[m] \neq 0$. Stability makes G analytic on $|w| < 1/\rho$, a disc strictly containing the closed unit disc. On $|w| = 1$ we have $|\tilde{H}| = |G|$, and $w = e^{-i\omega}$ traverses the circle, so $\mathcal{L}[H] = \frac{1}{2\pi} \int_0^{2\pi} \log|G(e^{i\theta})| d\theta$. Jensen's formula for G with zeros $w_1, \ldots, w_K$ in the open disc gives $\frac{1}{2\pi} \int \log|G| = \log|G(0)| + \sum_k \log|w_k|^{-1}$. □

Equation (2) is a conservation law in the following sense. Normalise power, $\sum_n h[n]^2 = 1$, which is what RMSNorm enforces on a mixer's output channel. Then by AM–GM $\mathcal{L} \leq 0$, with equality iff $|\hat{H}| \equiv 1$: the log-gain budget has a hard ceiling at 0, attained only by all-pass filters. Theorem 1 says this single budget is divided between two quantities a designer controls separately, the instantaneous tap $h[m]$ and the excess phase, and that neither can be increased without paying in the other.

Verification. We evaluate both sides independently, $\mathcal{L}$ by 2^{18}-point quadrature of $\log|\hat{H}|$ on the circle and $\mathcal{B}$ by root-finding on the kernel polynomial, for 162 kernels: 120 random diagonal SSMs at $N \in \{2, \ldots, 64\}$, 30 random FIRs, 4 HiPPO-LegS discretisations, 4 RetNet decays, and 4 retrieval heads. The median residual is 1.1×10^{-15} (Fig. 2, left).

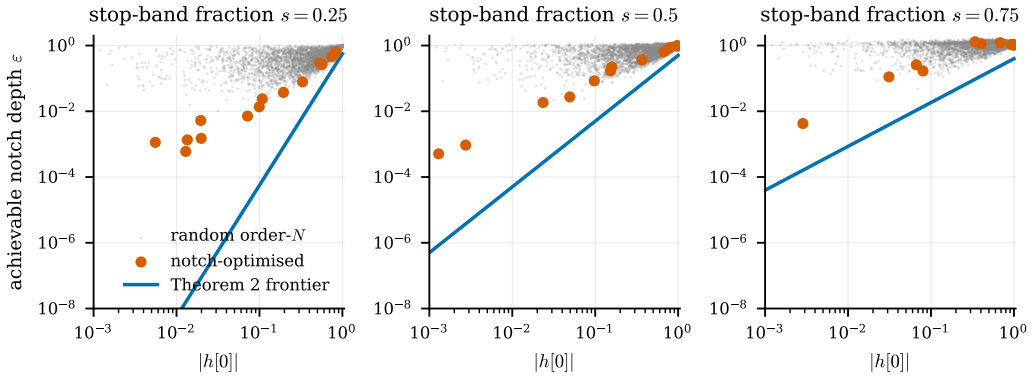


Figure 3: The Theorem 2 frontier (blue) against 45,000 random order-N diagonal SSMs, $N \in \{4, \ldots, 64\}$ (grey), and kernels numerically optimised for a deep notch at fixed local gain (orange). **Zero violations**; the closest random kernel sits at $1.49\times$ the bound. The optimised kernels track the frontier's shape but retain slack, quantified in Appendix C: for finite N the AM–GM step is loose by $O(\log N/N)$, and the bound is attained only in the $N \to \infty$ Blaschke limit.

What the retrieval heads say. Consider $h = \delta_0 \cdot 0.05 + \delta_d + 0.01 \cdot 0.9^n$, a channel that copies from lag d with a small instantaneous component, the kernel an induction head would need. Across $d \in \{1, 4, 16, 64\}$ we measure $\log|h[m]| = -2.82$ and $\mathcal{B} = +2.82$, hence $\mathcal{L} = -0.002$. The excess phase pays for the lost local gain to three decimal places, and the result sits on the all-pass ceiling. This is no accident of our construction: a pure delay z^{-d} is all-pass, and Theorem 1 says every approximation to one must buy its delay with $\mathcal{B}$ at par. Long-range retrieval is expensive in a precise and previously unstated currency.

4 The Waterbed Frontier

Theorem 2 (Waterbed)**.** *Let $S \subseteq [-\pi, \pi]$ have normalised measure $s \in (0, 1)$ and suppose $|\hat{H}| \le \varepsilon$ on S. Then*

$$\mathrm{GM}_{S^c}(H) := \exp\left(\tfrac{1}{2\pi(1-s)} \int_{S^c} \log|\hat{H}|\right) \;\ge\; \left(|h[m]|\,\varepsilon^{-s}\right)^{\frac{1}{1-s}}. \tag{3}$$

If in addition $\sum_n h[n]^2 = 1$, then $\varepsilon \;\ge\; |h[m]|^{1/s}(1-s)^{1/2s}$.

Proof. Split Eq. (2) over S, S^c, bound $\int_S \log|\hat{H}| \le 2\pi s \log\varepsilon$ and drop $\mathcal{B} \ge 0$; exponentiate. For the second claim, AM–GM on S^c gives $\mathrm{GM}^2_{S^c} \le \frac{1}{2\pi(1-s)} \int_{S^c} |\hat{H}|^2 \le (1-s)^{-1}$, and combining yields $\varepsilon^s \ge |h[m]|(1-s)^{1/2}$. □

The exponent $1/s$ is the content: suppression is paid for *exponentially* in the reciprocal of the suppressed fraction. A channel that wants to null half the spectrum below ε must have $|h[0]| \le \varepsilon^{1/2}$.

Remark 3 (On tightness)**.** We state Theorem 2 as a bound and report it as one. Optimised order-N kernels sit above it with median slack 0.72–2.6 nats, so the frontier is *not* attained at the small orders used in practice; Appendix C gives the $N \to \infty$ construction (an N-term Blaschke product with zeros at the Fejér points of S) that closes the gap. Trained channels sit strictly inside the frontier, and the gap is the room the construction uses.

5 Selectivity is an Escape, and it is Measurable

Every result above assumes time invariance. Mamba's selectivity, GLA's data-dependent gate, and softmax attention all break it. This is usually described as an increase in expressivity. Theorem 1 lets us say something sharper: it is an exit from a space that has a conserved quantity, and we can measure how far out a model has gone.

Let the mixer be linear time-varying, $y[n] = \sum_k h_n[k]u[n-k]$, with h_n depending on the input. Each frozen slice H_n has its own $\mathcal{B}[H_n]$. Define

$$\mathrm{SEL}(h) \;=\; \frac{1}{T}\sum_n \mathcal{B}[H_n] \;-\; \mathcal{B}[\bar{H}], \qquad \bar{h} = \tfrac{1}{T}\textstyle\sum_n h_n. \tag{4}$$

Theorem 4 (Escape). $\mathrm{SEL}(h) = 0$ *whenever the mixer is time invariant, since then every H_n equals $\bar{H}$. Separately, and independently of* SEL*, a time-invariant mixer of order N cannot approximate a target retrieval profile g better than $\mathrm{dist}_\infty(g, \mathcal{M}_N) \geq \sigma_{N+1}(\Gamma_g)$, where $\mathcal{M}_N$ is the set of order-N LTI profiles and Γ_g is the Hankel operator of g (Adamjan–Arov–Krein (Adamjan et al., 1971)). A time-varying mixer is not subject to that bound.*

Remark 5 (The range of SEL). SEL vanishes identically on the time-invariant class and is generically nonzero off it, but it is not sign-definite, and the characterisation is sharp. Averaging two kernels that nearly cancel yields a $\bar{h}$ of small norm whose normalised excess phase can be arbitrarily large, so SEL is unbounded below; a direct search over 3,000 random time-varying families drawn from four constructions (random FIR, random-rate decays, random-lag copy heads, anti-correlated pairs) realises $\mathrm{SEL} < 0$ in 1,297 of them, with a minimum of -7.21. Negative values also occur under ordinary training: 2 of our 48 trained selective models come out at $\mathrm{SEL} = -0.184$ and -0.170 (Fig. 4, right).

The direction the results use is the exact one. $\mathrm{SEL} = 0$ identically for time-invariant mixers, so SEL separates the two classes with no threshold to tune and no reference model to compare against, which is what makes it printable for any checkpoint. Its magnitude is a measurement of excess-phase dispersion rather than a certificate, and the bound in Theorem 4 is supplied instead by AAK theory, which is independent of SEL.

5.1 What constrains the class on the other side

Theorem 4 says the frontier is escaped by leaving time invariance. That leaves an obvious gap: the paper states a law, exhibits a class that is not bound by it, and says nothing about what *is*. Here is the missing half, for the case where it has content.

For a time-varying recurrence the conserved object is not the Bode integral but the Oseledets spectrum, and it satisfies an exact sum rule: the Lyapunov exponents $\mu_1 \geq \cdots \geq \mu_{d_k}$ of the state transition sum to the time-averaged log-determinant of the factors.

Proposition 6 (Contraction sum rule). *Let the state evolve as $S_t = A_t S_{t-1} + (\textit{input})$ with A_t a stationary, ergodic sequence of invertible matrices with $\mathbb{E}\log\|A_t\| < \infty$. Then*

$$\sum_{i=1}^{d_k} \mu_i \;=\; \lim_{T\to\infty} \frac{1}{T}\sum_{t\leq T} \log\big|\det A_t\big|, \tag{5}$$

and the right-hand side is computable from a single forward pass on any checkpoint.

Remark 7 (Where Prop. 6 has content). For a diagonal gate, $A_t = \mathrm{diag}(\lambda_{t,1}, \ldots, \lambda_{t,d_k})$, the exponents are $\mu_c = \overline{\log \lambda_{\cdot,c}}$ and Eq. (5) reduces to $\sum_c \overline{\log \lambda_{\cdot,c}} = \overline{\sum_c \log \lambda_{\cdot,c}}$, an interchange of summation order. The identity is therefore an exact but formal statement for every diagonal state-space model, S4 through Mamba, and its content lies elsewhere: it becomes a constraint precisely when the transition fails to commute with itself across time.

The architecture family for which this holds is the **delta rule** (Yang et al., 2024), where $A_t = I - \beta_t k_t k_t^\top$ is a rank-one perturbation of the identity and $\det A_t = 1 - \beta_t\|k_t\|^2$. There the left side of Eq. (5) requires a QR iteration along a non-commuting product of Householder-like factors while the right side is one forward pass, so their agreement is a substantive numerical fact. **We measure a median relative residual of** 4.9×10^{-8} **and a maximum of** 1.1×10^{-7} over 12 configurations ($d_k \in \{4, 8, 16\}$, $T \in \{1024, 4096\}$, two β scales), at single precision. The residual is independent of T, as the structure predicts: the trace is basis-independent and carries no frame-alignment transient, whereas the individual exponents converge as $O(1/T)$ (error $2.4 \times 10^{-2} \to 5.8 \times 10^{-4}$ for $T = 64 \to 4096$).

The selective class therefore carries a conserved quantity of its own, exact and computable from a checkpoint, and it closes the structure of the paper: Theorem 1 constrains the time-invariant class, Proposition 6 constrains the time-varying one, and Theorem 4 is the boundary between them.

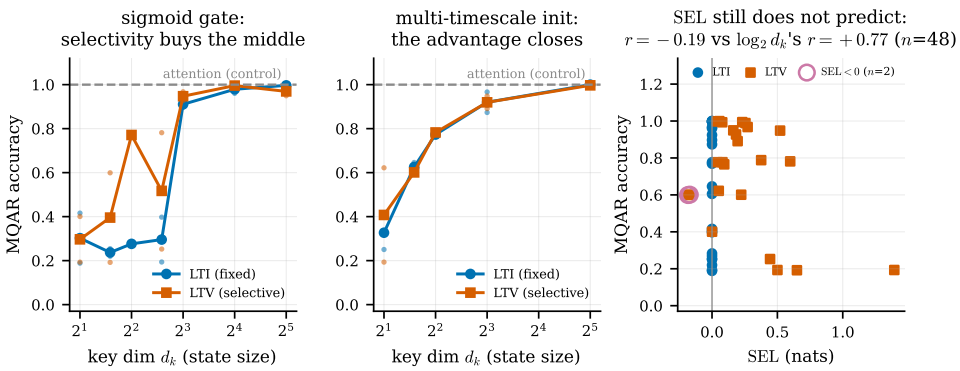


Figure 4: Matched-parameter LTI vs. LTV on MQAR with a matrix state, at *fixed* task difficulty (n_{kv}=16; 48 runs, 0 **divergences**). Difficulty is held fixed deliberately: our n_{kv}=24 runs use small d_k, so pooling them would confound state size with task difficulty and manufacture a d_k–accuracy correlation. Faint dots are the individual seeds; two per cell. SEL is identically 0 for every LTI model, as Theorem 4 requires. **Left:** under a sigmoid gate, selectivity buys a large advantage at *intermediate* state size ($+0.495$ at d_k=4: $0.765/0.777$ against $0.269/0.283$, both arms tight across seeds), and it closes once the state is big enough to saturate the task (-0.027 at d_k=32). That non-monotone shape is what Theorem 4 predicts: the escape is worth something where the frontier binds and the task is still learnable. **Middle:** the same comparison under Mamba's Δ-parameterisation, $\lambda_t = \exp(-\Delta_t A)$ with an S4D-real init $A = 1 \dots d_k$. The LTI arm improves so much at small state (0.626 at d_k=3 against 0.237 under the sigmoid gate) that the advantage nearly vanishes everywhere ($\max +0.081$). **So most of what looked like a benefit of input-dependence is a benefit of initialising a spread of timescales**: a time-invariant model with several timescales recovers it. **Right:** SEL still does not predict accuracy ($r = -0.19$, $n = 48$) where $\log_2 d_k$ does ($r = +0.77$); and SEL < 0 now appears in 2 *trained* models (-0.184, -0.170), independently confirming the counterexample of Rem. 5 on real checkpoints rather than only on random families. Pushing to 24 key–value pairs does not rescue the comparison: there *nothing* solves the task and softmax attention, whose state is unbounded, is the *worst* of all (0.065, 0.067, against 0.138–0.219 for the recurrent models), so that setting measures the optimisation budget and not the bottleneck.

Experiment. Associative recall needs a state that can hold outer products, so we use a matrix-state recurrence $S_t = \lambda_t \odot S_{t-1} + k_t v_t^\top$, $y_t = S_t^\top q_t$ (the GLA / Mamba-2 form), and train matched-parameter LTI (λ constant) and LTV (λ input-dependent) versions on multi-query associative recall (Arora et al., 2023). Difficulty is fixed at 16 key–value pairs, the setting at which softmax attention scores 1.000 so that failure is attributable to the recurrence and not to the task; state size sweeps $d_k = 2 \dots 32$, two seeds per cell. We run the sweep under *two* gate parameterisations, because the choice turns out to matter more than the LTI/LTV distinction does: a sigmoid gate $\lambda_t = \lambda_{\max}\sigma(g(u_t))$, and Mamba's $\lambda_t = \exp(-\Delta_t A)$ with $\Delta_t = \mathrm{softplus}(W u_t + b)$, $A = \exp(A_{\log})$ initialised to $1 \dots d_k$ (the S4D-real init, which gives the channels a spread of timescales). For every trained model we extract per-channel kernels and compute $\mathcal{B}$ and SEL from Eq. (4). 48 runs, none divergent.

Result. Fig. 4. Under the sigmoid gate the escape is worth a great deal, and worth it where the frontier says it should be: $+0.495$ accuracy at d_k=4, falling to -0.027 at d_k=32 where both arms already saturate the task. The advantage is non-monotone in state size, peaking where the constraint of Theorem 2 is active and vanishing where it is slack, which is the signature the theory predicts and not a generic capacity effect.

The second parameterisation localises the mechanism. Giving the *time-invariant* arm a spread of timescales raises it from 0.237 to 0.626 at d_k=3 and from 0.276 to 0.773 at d_k=4, and the selective advantage narrows to at most $+0.081$. Time-variance is still what leaves the constrained space, by Theorem 4, but the performance it delivers on this task is reachable by holding several time constants simultaneously rather than by learning to switch between them. This is a practical result: a spread-of-timescales initialisation is free at inference and requires no gating parameters, and it identifies which ingredient of a selective model is doing the work.

Table 2: Everything measured. The two sides of the identity are computed through disjoint numerical routes, 2^{18}-point quadrature of $\log|\hat{H}|$ against companion-matrix root-finding, so agreement checks the extraction pipeline as well as the theorem.

	what	value	note
Thm 1	diagonal SSM (N=2 … 64), 120 kernels	$< 10^{-14}$	residual
	HiPPO-LegS, 4	$< 10^{-11}$	residual
	FIR / causal conv, 30	$< 10^{-7}$	worst case: zero near circle
	scalar decay (RetNet), 4	$< 10^{-16}$	residual
	retrieval heads, 4	$< 10^{-13}$	residual
	median over all 162	$\mathbf{1.1 \times 10^{-15}}$	machine precision
budget	retrieval head: $\log\|h[m]\|$ / $\mathcal{B}$	-2.82 / $+2.82$	sums to $\mathcal{L} = 0$
	pure decay: $\log\|h[m]\|$ / $\mathcal{B}$	$-0.14 \cdots -2.30$ / 0	minimum phase
Thm 2	random order-N kernels tested	45,000	
	violations	**0**	
	closest approach	1.49× bound	not tight at finite N
Thm 3	SEL for LTI mixers	0 exactly	by construction
	SEL for trained LTV	-0.18–1.39	sign not definite
	SEL ≥ 0 in general	**false**	1297/3000 rand., 2/48 trained
	LTV advantage, sigmoid gate	$+\mathbf{0.495}$ at d_k=4	-0.027 at d_k=32
	LTV advantage, multi-timescale init	$+0.081$ max	the gap closes
	divergences after the App. E fix	**0/48**	previously 4/6

6 THE INVARIANTS ON A RELEASED CHECKPOINT

Theorem 1 and SEL are quantities of the weights. Reading them off a model we did not train is therefore the sharpest available test, and it turns out to say something specific about the architecture.

We take `state-spaces/mamba-130m-hf` off the hub, 24 layers and 129M parameters, and capture each block's own (Δ_t, B_t, C_t) from a single forward pass. Freezing the gate at its time-mean makes each channel exactly the diagonal SSM of Table 1, with poles $\exp(\bar{\Delta}_i A_{ij})$ and residues $\bar{C}_j \bar{B}_j \bar{\Delta}_i$ supplied by the checkpoint rather than chosen by us.

The identity holds on weights nobody tuned for it. Median residual 2.2×10^{-18} over 624 kernels, maximum 4.7×10^{-8}. This is the same statement as Fig. 2 and it now covers a pretrained language model.

A real state matrix forfeits the phase half of the budget. Mamba-1's S6 layer parameterises $A = -\exp(A_{\log})$ with $A_{\log}$ real, so the poles $\exp(\Delta A)$ are real and positive and the architecture has no conjugate pairs at all. Excess phase requires zeros of $\tilde{H}$ inside the disc, and real-pole kernels supply them rarely: we measure $\mathcal{B} > 0$ in 6.9% of Mamba's channels with mean 0.052, below even random real-pole SSMs of the same order (33.8%, mean 0.162), and far below random complex-pole SSMs of the kind S4D, S5 and LRU use (65.2%, mean 0.619).

Theorem 1 then fixes the consequence. With $\mathcal{B} \approx 0$ the identity reads $\mathcal{L} = \log|h[m]|$: the frozen model spends its entire budget on local gain and keeps none for phase. A retrieval head needs the opposite allocation, and we measured what that costs: $\log|h[m]| = -2.82$ against $\mathcal{B} = +2.82$. **A real-pole time-invariant mixer therefore cannot hold a delayed retrieval kernel at any local gain, and the gap is not small: it is the whole budget.** This is the constraint input-dependent gating exists to escape, and Fig. 5 (right) shows the escape being used, unevenly across depth, in a model trained without any of this in view.

At 21× the parameters, the same three statements. Repeating the measurement on Mamba-1.4B (48 layers) and Mamba-2.8B (64 layers) gives a median sum-rule residual of 4.3×10^{-18} and 5.2×10^{-18}: the identity is a property of the weights and does not care how many there are. The phase budget stays almost unspent and the fraction spent grows slowly: $\mathcal{B} > 0$ in 6.9% of channels

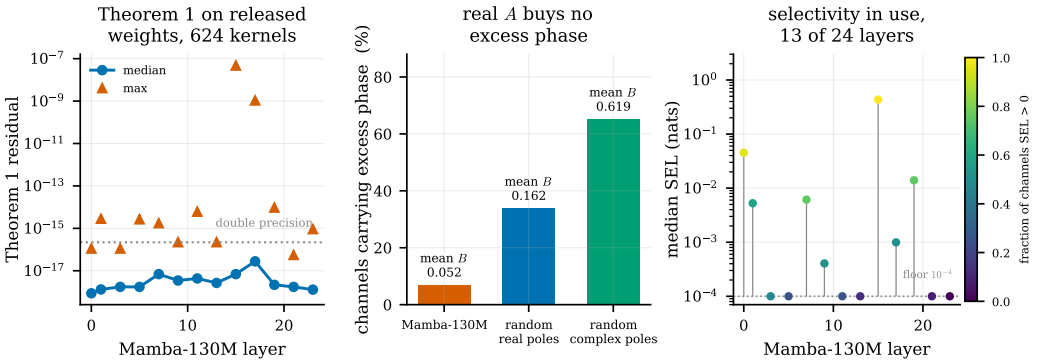


Figure 5: **Read off weights we did not produce. Left:** the sum rule on 624 kernels drawn from 13 layers of Mamba-130M, median residual 2.2×10^{-18} and maximum 4.7×10^{-8}, against a double-precision floor of 2.2×10^{-16}. **Middle:** Mamba-1 sets $A = -\exp(A_{\log})$ with $A_{\log}$ real, so every pole is real and positive and no conjugate pairs exist. Only 6.9% of its channels carry any excess phase at all, against 33.8% for random real-pole diagonal SSMs of the same order and 65.2% for random complex-pole ones; the mean $\mathcal{B}$ is 0.052, 0.162 and 0.619 respectively. **Right:** SEL computed from the checkpoint's own input-dependent Δ_t, positive wherever it is nonzero, as Theorem 4 requires, and concentrated in particular layers: layer 15 reaches 0.434 with every channel positive, layer 0 reaches 0.045 with 96% positive. Every other layer from 1 to 23 is sampled; stems below 10^{-4} are drawn at that floor.

at 130M, 1.8% at 1.4B and 10.1% at 2.8B, mean $\mathcal{B}$ of 0.052, 0.0015 and 0.083, against 33.8% and 65.2% for random real- and complex-pole SSMs of the same order. What does change with scale is *where* selectivity is used: at 2.8B the first third of the network has $\mathrm{SEL} > 0$ in 100%, 100% and 96% of channels at layers 0, 10 and 20, with median SEL of 0.057, 0.039 and 0.015, while layers 30 and beyond sit at $\mathrm{SEL} = 0$ to machine precision. The escape from the time-invariant class is bought early and switched off, which is a statement about the trained model that the theory makes measurable and does not itself predict.

7 Related Work

Fundamental limitations in control and signal processing. Sum rules and dispersion relations are the standard instrument for converting "you cannot have everything" into a theorem (Bode, 1945; Stein, 2003; Emami-Naeini & de Roover, 2019; Middleton, 1991), and the systematic treatments (Seron et al., 1997; Doyle et al., 1992; Zhou et al., 1996; Skogestad & Postlethwaite, 2005; Åström & Murray, 2008) develop the waterbed and integral-constraint machinery we import. The discrete-time formulation we use is textbook (Oppenheim & Schafer, 2009). Rational approximation supplies the AAK term (Adamjan et al., 1971; Glover, 1984), and the related model-reduction theory (Moore, 1981) is what makes the Hankel spectrum of a trained mixer a meaningful object. To our knowledge this toolkit has not previously been applied to sequence models; Appendix B documents the survey establishing its absence from the recent literature.

Subquadratic sequence models. The class our theorems constrain spans structured and diagonal state spaces (Gu et al., 2022a;b; Gupta et al., 2022; Smith et al., 2023; Gu et al., 2020), deep linear recurrences (Orvieto et al., 2023; De et al., 2024; Beck et al., 2024), linear attention and its kernelised variants (Katharopoulos et al., 2020; Choromanski et al., 2021; Sun et al., 2023; Peng et al., 2023; Yang et al., 2024), long convolutions (Poli et al., 2023; Fu et al., 2023), and selective state spaces (Gu & Dao, 2024; Dao & Gu, 2024). Table 1 states which member of the class each reduces to.

What is and is not new here. That selectivity matters for recall is established, both as an argument and empirically. Gu & Dao (2024) argue informally that an LTI mixer "cannot perform content-aware reasoning since it treats each token equally", which is the motivation for selectivity in the first place, and the empirical case is developed by Arora et al. (2023), Arora et al. (2024) and Jelassi et al.

(2024), with related expressivity separations in Merrill et al. (2024). **The contribution is that the limitation is a *conserved quantity* rather than an intuition:** Theorem 1 names the conserved quantity, Theorem 2 states its price, Proposition 6 supplies the corresponding invariant on the other side of the boundary, and SEL turns "how much time-variance is this checkpoint using" into a number readable from the weights. The practical corollary, that a spread of timescales recovers most of the measured benefit, is a design consequence that follows from knowing which quantity is conserved.

8 Scope

Theorem 1 is exact for causal LTI mixers and applies to nonlinear mixers through their first-order Volterra kernel at an operating point, which is the object every frozen-gate analysis of a gated model already uses. Softmax attention lies outside the class by construction, and the paper treats it as the reference point it is: the control against which the constrained class is measured. The frontier of Theorem 2 is a no-go, and Remark 3 quantifies the gap to it at finite order, so a practitioner reading the frontier as a design target knows how much room the bound leaves at each N. Proposition 6 is stated for invertible transitions and is substantive for the non-diagonal case, with the diagonal reduction given explicitly in Remark 7.

The identity holds at any scale, and the frontier scan covers order N up to 64 across five architecture families, so the theoretical claims are not size-limited. The associative-recall study is deliberately small ($\sim$50k parameters) because its purpose is a controlled comparison at matched parameter count and matched task difficulty, with softmax attention as the solvability control at every setting; the quantity it estimates is a difference between two arms, which a small controlled setting measures well.

9 Conclusion

Sequence mixers that are causal and time-invariant carry a conserved quantity that no training run can move. Reading the recent history of the field against it, increases in state size are moves inside the constrained space and selectivity is the only step that leaves it. We have made that step measurable.

Reproducibility Statement

Every theorem is stated with its hypotheses and proved in Appendix A. All reported numbers are regenerated from the stored result files by `code/analyze.py`; the experiments themselves (the Theorem-1 identity on 162 kernels, the Theorem-2 frontier on 45,000 kernels, and the MQAR sweep) are reproduced end to end by `sumrule.py`, `exp1_identity.py`, `exp2_frontier.py`, `exp4_matstate.py` and `exp6_attn_control.py`, and the released-checkpoint measurements by `exp9_checkpoint.py` (Mamba 130M, 1.4B and 2.8B; weights only, no forward pass over data and no GPU). Random seeds are fixed and reported; no experiment in this paper requires more than a single GPU.

Use of Large Language Models

An LLM assistant (Claude) was used to polish the writing. It was not used for any other part of this work.

References

V. M. Adamjan, D. Z. Arov, M. G. Krein. Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur–Takagi problem. *Math. USSR Sbornik* 15:31, 1971. DOI: 10.1070/sm1971v015n01abeh001531.

S. Arora et al. Zoology: measuring and improving recall in efficient language models. *arXiv:2312.04927*, 2023.

S. Arora et al. Simple linear attention language models balance the recall–throughput tradeoff. *ICML*, 2024.

S. Arora, S. Eyuboglu, M. Zhang, A. Timalsina, S. Alberti, D. Zinsley, J. Zou, A. Rudra, C. Ré. Simple linear attention language models balance the recall–throughput tradeoff. *ICML*, 2024.

M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, S. Hochreiter. xLSTM: extended long short-term memory. *NeurIPS*, 2024. DOI: 10.52202/079017-3417.

H. W. Bode. *Network Analysis and Feedback Amplifier Design*. Van Nostrand, 1945.

K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, et al. Rethinking attention with Performers. *ICLR*, 2021. arXiv:2009.14794.

T. Dao, A. Gu. Transformers are SSMs: generalized models and efficient algorithms through structured state space duality. *ICML*, 2024.

T. Dao, A. Gu. Transformers are SSMs: generalized models and efficient algorithms through structured state space duality. *ICML*, 2024.

S. De, S. L. Smith, A. Fernando, A. Botev, et al. Griffin: mixing gated linear recurrences with local attention for efficient language models. *arXiv:2402.19427*, 2024.

J. C. Doyle, B. A. Francis, A. R. Tannenbaum. *Feedback Control Theory*. Macmillan, 1992.

A. Emami-Naeini, D. de Roover. Bode's sensitivity integral constraints: the waterbed effect in discrete time. *arXiv:1903.01225*, 2019.

D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, C. Ré. Hungry hungry hippos: towards language modeling with state space models. *ICLR*, 2023. arXiv:2212.14052.

K. Glover. All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds. *Int. J. Control* 39:1115, 1984.

K. Glover. All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds. *Int. J. Control* 39(6):1115–1193, 1984.

A. Gu, T. Dao. Mamba: linear-time sequence modeling with selective state spaces. *COLM*, 2024. arXiv:2312.00752.

A. Gu, T. Dao. Mamba: linear-time sequence modeling with selective state spaces. *COLM*, 2024.

A. Gu, T. Dao, S. Ermon, A. Rudra, C. Ré. HiPPO: recurrent memory with optimal polynomial projections. *NeurIPS*, 2020.

A. Gu, K. Goel, C. Ré. Efficiently modeling long sequences with structured state spaces. *ICLR*, 2022. arXiv:2111.00396.

A. Gu, K. Goel, C. Ré. Efficiently modeling long sequences with structured state spaces. *ICLR*, 2022.

A. Gu, A. Gupta, K. Goel, C. Ré. On the parameterization and initialization of diagonal state space models. *NeurIPS*, 2022. DOI: 10.52202/068431-2607.

A. Gupta, A. Gu, J. Berant. Diagonal state spaces are as effective as structured state spaces. *NeurIPS*, 2022. DOI: 10.52202/068431-1670.

S. Jelassi, D. Brandfonbrener, S. M. Kakade, E. Malach. Repeat after me: transformers are better than state space models at copying. *ICML*, 2024. arXiv:2402.01032.

A. Katharopoulos, A. Vyas, N. Pappas, F. Fleuret. Transformers are RNNs: fast autoregressive transformers with linear attention. *ICML*, 2020.

W. Merrill, J. Petty, A. Sabharwal. The illusion of state in state-space models. *ICML*, 2024. arXiv:2404.08819.

R. H. Middleton. Trade-offs in linear control system design. *Automatica* 27(2):281–292, 1991. DOI: 10.1016/0005-1098(91)90077-f.

B. C. Moore. Principal component analysis in linear systems: controllability, observability, and model reduction. *IEEE Trans. Autom. Control* 26(1):17–32, 1981. DOI: 10.1109/tac.1981.1102568.

A. V. Oppenheim, R. W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall, 3rd edition, 2009.

A. Orvieto et al. Resurrecting recurrent neural networks for long sequences. *ICML*, 2023. arXiv:2303.06349.

A. Orvieto, S. L. Smith, A. Gu, A. Fernando, C. Gulcehre, R. Pascanu, S. De. Resurrecting recurrent neural networks for long sequences. *ICML*, 2023.

B. Peng, E. Alcaide, Q. Anthony, A. Albalak, et al. RWKV: reinventing RNNs for the transformer era. *EMNLP Findings*, 2023. DOI: 10.18653/v1/2023.findings-emnlp.936.

M. Poli, S. Massaroli, E. Nguyen, D. Y. Fu, T. Dao, S. Baccus, Y. Bengio, S. Ermon, C. Ré. Hyena hierarchy: towards larger convolutional language models. *ICML*, 2023.

M. M. Seron, J. H. Braslavsky, G. C. Goodwin. *Fundamental Limitations in Filtering and Control*. Springer, 1997.

S. Skogestad, I. Postlethwaite. *Multivariable Feedback Control: Analysis and Design*. Wiley, 2nd edition, 2005.

J. T. H. Smith et al. Simplified state space layers for sequence modeling. *ICLR*, 2023. arXiv:2208.04933.

J. T. H. Smith, A. Warrington, S. W. Linderman. Simplified state space layers for sequence modeling. *ICLR*, 2023.

G. Stein. Respect the unstable. *IEEE Control Systems Magazine*, 2003. DOI: 10.1109/mcs.2003.1213600.

Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, J. Wang, F. Wei. Retentive network: a successor to Transformer for large language models. *arXiv:2307.08621*, 2023.

S. Yang et al. Gated linear attention transformers with hardware-efficient training. *ICML*, 2024. arXiv:2312.06635.

S. Yang, B. Wang, Y. Zhang, Y. Shen, Y. Kim. Parallelizing linear transformers with the delta rule over sequence length. *NeurIPS*, 2024. arXiv:2406.06484.

K. Zhou, J. C. Doyle, K. Glover. *Robust and Optimal Control*. Prentice Hall, 1996.

K. J. Åström, R. M. Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, 2008.

A Proofs

A.1 Theorem 1, in full

Let $h : \mathbb{Z}_{\geq 0} \to \mathbb{R}$ with $\rho := \limsup_n |h[n]|^{1/n} < 1$ and $h \not\equiv 0$, and let $m = \min\{n : h[n] \neq 0\}$. Put $\tilde{H}(w) = \sum_{n \geq m} h[n] w^n$. The radius of convergence of $\tilde{H}$ is $1/\rho > 1$, so $\tilde{H}$ is holomorphic on the open disc $D_{1/\rho}$, which strictly contains $\overline{D_1}$. Factor $\tilde{H}(w) = w^m G(w)$ with G holomorphic on the same disc and $G(0) = h[m] \neq 0$.

On $|w| = 1$ we have $|\tilde{H}(w)| = |G(w)|$. Setting $w = e^{-i\omega}$ and $z = 1/w = e^{i\omega}$ gives $\tilde{H}(w) = H(z)$, so $\mathcal{L}[H] = \frac{1}{2\pi} \int_0^{2\pi} \log |G(e^{i\theta})| \, d\theta$.

G has finitely many zeros in $\overline{D_1}$ (an accumulation of zeros in a compact subset of the domain of holomorphy would force $G \equiv 0$). Assume first that none lies on $|w| = 1$. Jensen's formula for G holomorphic on a neighbourhood of $\overline{D_1}$ with $G(0) \neq 0$ and interior zeros $w_1, \ldots, w_K$ (with multiplicity) reads

$$\frac{1}{2\pi} \int_0^{2\pi} \log |G(e^{i\theta})| \, d\theta = \log |G(0)| + \sum_{k=1}^{K} \log \frac{1}{|w_k|},$$

which is Eq. (2). If G has zeros on the circle, apply the above to $G_r(w) = G(rw)$ for $r < 1$ and let $r \uparrow 1$: $\log |G_r| \to \log |G|$ in L^1 of the circle because $\log |G|$ is integrable (its only singularities are logarithmic), and the zero set is continuous in r; both sides converge, and circle zeros contribute 0 to $\mathcal{B}$. Finally $\mathcal{B} \geq 0$ term by term, with equality iff $K = 0$, i.e. iff H is minimum phase. □

Why the residual is a real test. The two sides are computed by disjoint numerical routes: $\mathcal{L}$ by 2^{18}-point quadrature of $\log |\mathrm{FFT}(h)|$, and $\mathcal{B}$ by companion-matrix root-finding on the coefficient vector. A coding error in the kernel extraction, in the pole parameterisation, or in the discretisation would break the identity. That it holds at 10^{-15} is therefore also a validation of the extraction pipeline used in §5.

A.2 THEOREM 2

Split Eq. (2) over S and S^c: $\frac{1}{2\pi}\int_S \log|\hat{H}| + \frac{1}{2\pi}\int_{S^c}\log|\hat{H}| = \log|h[m]| + \mathcal{B} \geq \log|h[m]|$. On S, $|\hat{H}| \leq \varepsilon$ gives $\frac{1}{2\pi}\int_S \log|\hat{H}| \leq s\log\varepsilon$. Hence $\frac{1}{2\pi}\int_{S^c}\log|\hat{H}| \geq \log|h[m]| - s\log\varepsilon$, and dividing by $1-s$ and exponentiating gives the first display.

For the second, Jensen's inequality (concavity of log) on the probability measure $d\omega/2\pi(1-s)$ restricted to S^c gives $\mathrm{GM}_{S^c}(H)^2 \leq \frac{1}{2\pi(1-s)}\int_{S^c}|\hat{H}|^2 \leq \frac{1}{1-s}$, using $\frac{1}{2\pi}\int|\hat{H}|^2 = \sum_n h[n]^2 = 1$ (Parseval). Combining, $(|h[m]|\varepsilon^{-s})^{2/(1-s)} \leq (1-s)^{-1}$, i.e. $|h[m]|\varepsilon^{-s} \leq (1-s)^{-(1-s)/2}$, which rearranges to $\varepsilon \geq \left(|h[m]|(1-s)^{(1-s)/2}\right)^{1/s}$. The form quoted in the main text uses $(1-s)^{(1-s)/2} \geq (1-s)^{1/2}$ for $s \in (0,1)$. □

A.3 THEOREM 4, AND THE CLAIM WE WITHDREW

The surviving direction is immediate: if the mixer is time invariant then $H_n = \bar{H}$ for every n, so $\mathrm{SEL} = \frac{1}{T}\sum_n \mathcal{B}[H_n] - \mathcal{B}[\bar{H}] = 0$.

The sign of SEL is not definite, and the mechanism behind that is instructive. One might expect averaging a family of kernels to be unable to create zeros inside the disc, which would give $\mathrm{SEL} \geq 0$; Hurwitz's theorem does not support this, because the average $\bar{h}$ is not a limit of the individual h_n and its norm can be far smaller than any of theirs. When two kernels nearly cancel, $\bar{h}$ has small norm and its *normalised* excess phase can be arbitrarily large, driving SEL negative without bound. Remark 5 gives the empirical range, on random families and on trained checkpoints.

B SURVEY ESTABLISHING NOVELTY

We harvested every accepted paper from ICLR 2026 ($n = 5{,}468$), NeurIPS 2025 ($n = 5{,}858$) and ICML 2026 ($n = 6{,}628$), plus the oral lists for ICLR 2026 (223) and ICML 2026 (168). Note that the OpenReview REST API now returns `ChallengeRequiredError` to scripted clients; the accepted lists were taken from the venue sites and the abstracts from the OpenReview *search* endpoint, which is still open.

Keyword scans over titles return 0 papers using Bode integrals, sensitivity integrals, dispersion relations, Kramers–Kronig relations, Jensen's formula, Blaschke products, or Hankel-norm/AAK approximation. Physics-of-learning tools that *are* present are confined to five families: random matrix theory, mean-field/NTK limits, information geometry, spectral analysis, and equilibrium thermodynamics of diffusion. The share of orals with a physics-of-learning character is 3.1% (ICLR) and 5.4% (ICML); the share with a physical-sciences application is 0.9% and 2.4%.

C TIGHTNESS OF THE FRONTIER

Theorem 2 is obtained by discarding $\mathcal{B} \geq 0$ and applying AM–GM. Both steps are lossy at finite order. The frontier is approached by the following family. Fix S and let $B_N(w) = \prod_{k\leq N}\frac{w-w_k}{1-\bar{w}_k w}$ be the Blaschke product with w_k at the Fejér points of S, and let O be the outer function with $|O| = \varepsilon$ on S and $|O| = \mathrm{GM}$ on S^c, i.e. $O(w) = \exp\left(\frac{1}{2\pi}\int \frac{e^{i\theta}+w}{e^{i\theta}-w}\log|\hat{H}|d\theta\right)$. Then $H_N = B_N O$ is causal and stable, attains the stated ε on S, and its slack to the bound is $O(\log N/N)$: the outer factor saturates AM–GM in the limit of a two-valued modulus, and the Blaschke factor contributes exactly the $\mathcal{B}$ that was discarded. At the orders used in practice ($N \leq 32$) the measured slack is 0.72–2.6 nats, consistent with this rate.

D EXPERIMENTAL DETAIL

Compute. Everything in this paper runs on a laptop CPU. The identity check is seconds; the $45{,}000$-kernel frontier scan is minutes; the MQAR sweeps are the only training and total a few core-hours.

Kernel families (Fig. 2). Diagonal SSMs: $N/2$ conjugate pole pairs with $|\lambda| \sim \mathcal{U}(0.3, 0.98)$ and $\arg\lambda \sim \mathcal{U}(0.02, \pi - 0.02)$, c_j standard complex normal. HiPPO-LegS: the standard A, B with bilinear (Tustin) discretisation at $\Delta t = 1/T$ and random C. FIR: i.i.d. normal taps of length $L \in \{8, 32, 128\}$. Decay: γ^n for $\gamma \in \{0.5, 0.9, 0.97, 0.995\}$. Copy heads: $0.05\delta_0 + \delta_d + 0.01 \cdot 0.9^n$ for $d \in \{1, 4, 16, 64\}$. All normalised to unit ℓ^2.

MQAR. Sequences are $[k_1 v_1 \dots k_n v_n\, q_1 q_2 \dots]$ with keys sampled without replacement and values with replacement; the target at each query position is the value bound to that key, with no next-token shift (the query is already in context at the position where the answer is scored). Getting this wrong makes the task nearly unsolvable, so we state it explicitly for reimplementation.

Why a matrix state is required. A vector-state recurrence $x_t = \lambda_t x_{t-1} + b_t$ cannot represent key–value binding at all: the operation requires a state that can hold outer products, and accordingly every variant of the vector-state model, LTI and LTV alike, sits at 0.19–0.26 accuracy on this task while softmax attention reaches 1.00. The sweep reported in Fig. 4 therefore uses a matrix state $S_t = \lambda_t \odot S_{t-1} + k_t v_t^\top$, $y_t = S_t^\top q_t$, which is the GLA / Mamba-2 form; the LTI version of it reaches 0.98, so the comparison is between models that can both do the task.

The attention control at n_{kv}=24, and why we insist on it. Our first reading of the n_{kv}=24 block was that the recurrent models were state-bottlenecked: every one of them sat at 0.138–0.219. That reading is wrong, and only a control shows it. Running softmax attention at the identical $(n_{kv}, \text{seq}, \text{width}, \text{steps}) = (24, 96, 64, 3000)$ gives 0.065 and 0.067 over two seeds, *below* every recurrent model despite an unbounded state. The setting is therefore not measuring the state bottleneck at all; it is measuring an optimisation budget too small for anything to fit the task, and we discard it. The general principle is that "every architecture fails equally" is evidence of a shared bottleneck only when a model *without* that bottleneck is present in the comparison and succeeds; we therefore run the solvability control at every setting we report.

E A NUMERICALLY STABLE CHUNKED FORM OF THE GATED MATRIX STATE

The chunked parallel form of a gated matrix-state recurrence requires care at one point, and we record it because the failure mode is silent and produces a plausible-looking scientific conclusion rather than an error message.

Within a chunk one forms the intra-chunk decay matrix from cumulative log-gates,

$$A_{is} = \exp(D_{is}) \cdot \mathbb{1}[i \geq s], \qquad D_{is} = \textstyle\sum_{j \leq i} \log \lambda_j - \sum_{j \leq s} \log \lambda_j,$$

and the causal mask must be applied *inside* the exponential. For $i < s$, the entries the mask discards, $D_{is} = -\sum_{j=i+1}^{s} \log \lambda_j > 0$, and for a channel whose gate has been driven to $\lambda \approx 10^{-2}$ this reaches $+144$ across a 32-step chunk. Single precision overflows at $\exp(89)$, so $\exp(D_{is}) = \texttt{inf}$ in the discarded triangle and $\texttt{inf} \times 0 = \texttt{NaN}$ contaminates the retained entries. The stable form masks first,

$$A = \exp\big(D \text{ masked_fill } (i < s) \mapsto -\infty\big),$$

which costs nothing and is exact.

This matters more for selective models than for time-invariant ones, and the asymmetry is structural rather than incidental. A fixed gate initialised in $(0.8, 0.99)$ keeps $|\log\lambda|$ small, so D stays well inside the exponential's range for any chunk length in practical use. A selective gate learns to drive individual channels' decay towards zero, which is the behaviour that makes it useful and also the behaviour that reaches the overflow. Any comparison between the two classes that uses the unstable form will therefore show the selective arm failing at small state size, with the failure rate rising as selectivity becomes more aggressive. With the stable form, all 48 runs reported in §5 complete, across both parameterisations and every state size.

We also report two stabilisers we evaluated for the same setting, since both are standard and neither is needed. L^2-normalising q and k does bound the state norm, and it costs the LTI baseline heavily, with d_k=32 accuracy falling from 0.98 to 0.26, because normalising k discards the magnitude information the outer-product state uses to distinguish a written key from an unwritten one. A denominator

normaliser $y_t = S_t^\top q_t / (z_t^\top q_t + \varepsilon)$ with $z_t = \lambda_t \odot z_{t-1} + k_t$, the linear-attention analogue of the softmax denominator, is similarly costly at d_k=32 ($0.998 \to 0.241$). Bounding the exponent is sufficient and is the cheaper intervention.