WHEN DO SPIKES MATTER? TIMING BEYOND RATES

Anonymous authors
Paper under double-blind review

ABSTRACT

When does a spiking neural network (SNN) depend only on how many spikes each neuron fires, and when does it also depend on when they fire? This matters: an SNN that ignores timing is only a time-stepped rate network, and one that uses timing loses information when spikes are reduced to counts. We study this question in wide networks of leaky integrate-and-fire (LIF) neurons with Gaussian weights, which pass on only the mean and the temporal second moment of each layer's spike trains. In wide networks this yields a kernel over spike trains, the spiking analogue of the neural-network Gaussian-process (NNGP) kernel, and rate coding keeps only its time-averaged part. It suffices for exact rate computation that synaptic filtering removes every temporal component except the average, and for approximate rate computation that little passes through. Conversely, we construct two spike populations with identical spike counts and firing rates for which one spiking layer produces different outputs, even when only their rates are read out. In this model timing matters only when structure survives filtering and reaches the readout; for trained networks these quantities are diagnostics.

1 INTRODUCTION

Spiking neural networks (SNNs) are receiving increasing attention because event-driven communication promises efficient inference, while spike timing offers a way to represent temporal structure. Deep SNNs are built mainly through two routes. ANN-to-SNN conversion maps a trained artificial neural network to spikes whose counts reproduce its activations, often retaining its accuracy at inference (Rueckauer et al., 2017; Bu et al., 2022). Direct training from scratch instead optimizes the spiking dynamics themselves, allowing learning to use when spikes occur rather than only how many occur (Yu et al., 2026). Conversion therefore keeps the rate computation of the source network. Direct training may learn a temporal computation, but nothing forces it to.

Which of the two a given network does cannot be read off its accuracy or its spikes. A converted network may emit irregular spike times that its output ignores, whereas a directly trained network may carry task information in those times, and the two spike rasters can look equally rich. The difference matters in practice: if timing is unused, the time steps of a simulation serve only to count spikes; if timing is used, any rate-based shortcut, such as conversion from a rate network or averaging spikes over time, discards information that the task needs. Whether a given SNN depends only on spike counts, or also on spike timing, therefore remains an open question.

Existing results establish both possibilities but do not connect them in one test. Conversion methods show that spike counts can reproduce rectified linear activations (Rueckauer et al., 2017; Bu et al., 2022), and an exact time-to-first-spike mapping shows that a timing code can still implement the same rectified linear function (Stanojevic et al., 2023). Conversely, constructed SNNs can exploit spike timing beyond counts (Maass, 1997), and directly trained networks can become sensitive to changes in spike timing that preserve counts (Yu et al., 2026). These results do not determine, for a given architecture, parameters and output readout, which information its function actually uses.

At the simplest level, two spike patterns can have equal counts and thus be identical after rate reduction, while stateful temporal dynamics preserve their ordering and can produce different readouts. We make this comparison in the function space induced by the network and its readout, as Figure 1 illustrates. For wide hard leaky integrate-and-fire networks with Gaussian weights, we derive the temporal kernel, the spiking counterpart of the neural-network Gaussian-process (NNGP) kernel of wide conventional networks, identify spike-count reduction with its constant (time-averaged) component, and test whether the output kernel changes when the remaining timing information is

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.

removed. For Gaussian networks this is a rigorous test; for trained finite networks the same decomposition gives forward-pass diagnostics.

The answer is conditional rather than uniformly yes or no. Timing matters only when temporal structure survives the synaptic dynamics and affects the readout; otherwise the network reduces to a rate computation. We prove exact and approximate sufficient conditions for this reduction, construct equal-count inputs that a single layer separates, and test when depth contracts their difference. Converted checkpoints from vision, language, speech, control and graphs recover their rate networks; in controlled comparisons, converted and rate-trained networks ignore rate-preserving timing changes, whereas networks trained on time-to-first-spike and burst codes fail when those times are permuted. We do not claim that direct training beats conversion. We explain why conversion can keep source-network accuracy without an SNN-specific temporal representation, and isolate the next learning question: when does training align spike timing with the task? *Our contributions are as follows:*

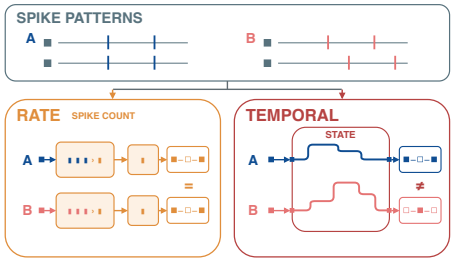


Figure 1: **Same rate, different computation.**

i. We derive a temporal Gaussian-process limit for wide hard-spiking networks and treat the threshold exactly, without a forward-pass surrogate.

ii. We distinguish three levels of rate reduction and give sufficient conditions for a network, including the conversion integrator, to factor through counts exactly or approximately.

iii. We construct equal-count spike populations that one hard-spiking layer separates, predict the direction of the separation in closed form, and test its local depth contraction.

iv. We connect this representation-level distinction to the readout and use the resulting quantities as diagnostics on converted and directly trained networks.

2 Related Work

Conversion and equivalence. ANN-to-SNN conversion methods map trained artificial neural networks to spiking networks. Threshold-balanced integrate-and-fire conversion approximates rectified linear activations with spike counts over a long horizon (Rueckauer et al., 2017), while short horizons introduce conversion error (Bu et al., 2022). An exact time-to-first-spike mapping fires one spike per neuron without loss of performance (Stanojevic et al., 2023), showing that temporal coding can implement a rectified linear network. These methods construct rate-equivalent SNNs but do not determine whether a separately trained network uses spike timing beyond spike counts.

Expressivity separations. The complementary direction shows that spiking networks can exceed rate networks. Networks of spiking neurons with temporal coding compute certain functions with fewer units than any sigmoidal network, and can simulate threshold circuits and Turing machines (Maass, 1996; 1997). These are existence results: a separating construction at one parameter setting does not determine whether another network uses temporal information; our Gaussian analysis instead asks which temporal information reaches a specified readout.

Neural coding. Neural-coding studies formulate the distinction in terms of temporal resolution and downstream readout. Temporal encoding is defined relative to an encoding time window (Theunissen & Miller, 1995), and the timescales at which spike times carry information form a spectrum rather than a dichotomy (Panzeri et al., 2026). Information present in a spike train need not be used by a downstream stage (Panzeri et al., 2017), and the rate-versus-spike debate concerns the causal role of spike timing rather than spike-train observables alone (Brette, 2015). These definitions apply to measured spike trains but do not provide a function-space criterion for a given network; we give one in the Gaussian model together with its temporal-resolution index.

Wide-network limits and propagation. Our technical setting is the Gaussian-process limit of wide networks (Lee et al., 2018) and the analysis of signal propagation with depth (Poole et al., 2016; Schoenholz et al., 2017). Both were developed for continuous activations and for a state that is a scalar covariance, whereas a stateful spiking neuron maps an entire Gaussian path to an entire spike

path, so the recursion must act on temporal matrices and the threshold cannot be smoothed away. On the dynamical-systems side, exact mean-field reductions of spiking populations close only when a synchrony variable is carried alongside the rate (Montbrió et al., 2015), the continuous-time counterpart of the second-order object our analysis finds to be irreducible. Empirical work perturbing the timing of trained networks (Yu et al., 2026) is complementary to ours: it probes trained finite networks, where our theorems describe the random-weight prior and supply diagnostics.

3 From Spike Populations to a Rate Criterion

A Gaussian synaptic layer sees the preceding spike population only through its mean path and temporal second moment (Section 3.2); in the wide limit this pair evolves deterministically and defines a temporal NNGP kernel (Section 3.3). Rate coding keeps only the constant temporal component of this state, so the question becomes whether the recursion factors through that component. We show when it does, exactly or approximately (Section 3.4), when it does not (Section 3.5), and when the difference reaches the readout (Section 3.5).

3.1 Setup: spiking layers, descriptors and the rate quotient

Fix inputs $\mathcal{X}_M = \{x^1, \dots, x^M\}$, time horizon T, and depth L. Neuron i in layer ℓ emits $\mathbf{S}_i^\ell(x) \in \{0,1\}^T$. A temporal synaptic operator $\mathbf{A}_\ell \in \mathbb{R}^{T\times T}$ acts before random mixing; a causal filter has $(\mathbf{A}_\ell \mathbf{s})_t = \sum_{q\le t} \kappa_\ell(t-q) s_q$. The drive and weight parameterization are

$$\mathbf{H}_i^{\ell+1}(x) = \mathbf{b}_{\ell+1}(x) + \sum_{j=1}^{n_\ell} W_{ij}^{\ell+1} \mathbf{A}_\ell \mathbf{S}_j^\ell(x), \qquad W_{ij}^{\ell+1} = \frac{\mu_{\ell+1}}{n_\ell} + \frac{\sigma_{\ell+1}}{\sqrt{n_\ell}} G_{ij}^{\ell+1}. \tag{1}$$

The $G_{ij}^{\ell+1}$ are iid standard Gaussian, independent between layers and of the preceding population; bias paths are deterministic and common across neurons, and the same weight acts at all times and on all inputs. The two scalings propagate population means and nonvanishing fluctuations, respectively.

For a drive $\mathbf{h}$, hard LIF evolves as

$$u_t = \lambda u_{t-1} + h_t - \rho s_{t-1}, \qquad s_t = \mathbf{1}\{u_t \ge \vartheta\}, \qquad t = 1, \dots, T. \tag{2}$$

Here $\lambda \in [0,1]$, $\rho \ge 0$, the initial membrane value u_0 is fixed, and $s_0 = 0$. Write $\mathbf{s} = \Psi_{\mathrm{LIF}}(\mathbf{h})$ and let $\Psi^{(M)}$ apply this map blockwise to the M inputs. Layer-specific neuron parameters are allowed.

Stack the input paths into $\mathbf{S}_i^\ell \in \{0,1\}^D$, where $D = MT$, and put $\mathbf{B}_\ell = \mathbf{I}_M \otimes \mathbf{A}_\ell$. The empirical descriptor and its temporal-kernel blocks are $\mathbf{m}_n^\ell = \frac{1}{n_\ell}\sum_i \mathbf{S}_i^\ell$, $\mathbf{Q}_n^\ell = \frac{1}{n_\ell}\sum_i \mathbf{S}_i^\ell (\mathbf{S}_i^\ell)^\top$ and $\mathbf{K}^\ell(x^a, x^b) = \mathbf{Q}_{ab}^\ell$. $\mathbf{Q}$ is an uncentered second moment, not a covariance; its diagonal equals $\mathbf{m}$, but keeping both makes the mean and fluctuation channels explicit. A realizable descriptor arises from a probability law on $\{0,1\}^D$; these descriptors form a compact set $\mathfrak{C}_D$.

The firing rate is $R(\mathbf{s}) = T^{-1}\mathbf{1}_T^\top \mathbf{s}$. To keep it distinct from the membrane value u_0, we write the normalized constant temporal vector as $\mathbf{v}_0 = T^{-1/2}\mathbf{1}_T$, and set $\mathbf{P}_0 = \mathbf{v}_0\mathbf{v}_0^\top$, $\mathbf{P}_\perp = \mathbf{I}_T - \mathbf{P}_0$, $\mathbf{\Pi} = \mathbf{I}_M \otimes \mathbf{P}_0$, and $\mathbf{\Pi}_\perp = \mathbf{I}_D - \mathbf{\Pi}$. The rate quotient is $\mathcal{R}(\mathbf{m}, \mathbf{Q}) = (\mathbf{\Pi}\mathbf{m}, \mathbf{\Pi}\mathbf{Q}\mathbf{\Pi})$ and $K_{\mathrm{rate}}(x, x') = \frac{1}{T}\mathbf{v}_0^\top \mathbf{K}(x, x')\mathbf{v}_0$. Equality of the joint rate law implies equality of this quotient, which retains only first and second rate moments, including cross-input blocks. Everything that follows compares the network's evolution of $(\mathbf{m}, \mathbf{Q})$ with its evolution of $\mathcal{R}(\mathbf{m}, \mathbf{Q})$.

3.2 Gaussian mixing transmits only the descriptor

The first step makes $(\mathbf{m}, \mathbf{Q})$ the state of the network, already at finite width.

Theorem 3.1 (Exact finite-width sufficiency)**.** *Let $\mathcal{F}_\ell$ be generated by the preceding spike population. Under* (1)*, the stacked drives are conditionally iid with law*

$$\mathbf{H}_i^{\ell+1} \mid \mathcal{F}_\ell \sim \mathcal{N}\big(\mathbf{b}_{\ell+1} + \mu_{\ell+1}\mathbf{B}_\ell \mathbf{m}_n^\ell,\ \sigma_{\ell+1}^2 \mathbf{B}_\ell \mathbf{Q}_n^\ell \mathbf{B}_\ell^\top\big)\,. \tag{3}$$

Consequently, the conditional law of the entire next spike population depends on the preceding population only through $(\mathbf{m}_n^\ell, \mathbf{Q}_n^\ell)$.

Conditioned on the paths, each drive is a sum of independent Gaussian coefficients times deterministic temporal vectors, whose outer products sum to $n_\ell \mathbf{Q}_n^\ell$; independent weight rows give conditional independence (Appendix A.1.2).

This is exact channel sufficiency, not moment closure: populations with the same descriptor may differ in higher-order path statistics yet are indistinguishable to the next layer. The descriptor need not be minimal, since filtering can discard further directions, which is precisely the mechanism of rate collapse in Section 3.4; nor is the finite-width *unconditional* drive Gaussian, being a mixture over random descriptors. The uncentered $\mathbf{Q}$, rather than $\mathbf{Q} - \mathbf{m}\mathbf{m}^\top$, is essential: synaptic fluctuations multiply the spike vectors themselves, so even deterministic spike paths produce random drives.

For a descriptor $c = (\mathbf{m}, \mathbf{Q})$, define $\mathbf{Z}_{\ell+1}(c)$ by the Gaussian law in (3) with empirical moments replaced by c. The exact output moment map is $\mathcal{T}_\ell(c) = (\mathbb{E}\Psi^{(M)}(\mathbf{Z}_{\ell+1}(c)),\ \mathbb{E}[\Psi^{(M)}(\mathbf{Z}_{\ell+1}(c))\Psi^{(M)}(\mathbf{Z}_{\ell+1}(c))^\top])$. It retains all within- and cross-input temporal dependence needed by the next layer.

3.3 The wide limit: a temporal NNGP recursion

As width grows, the empirical descriptor concentrates on its conditional mean $\mathcal{T}_\ell(c_n^\ell)$; passing to the limit also requires continuity of $\mathcal{T}_\ell$ despite the hard threshold. For continuity, fix a candidate spike pattern $\mathbf{s}$. Unrolling (2) expresses each threshold condition as an affine inequality in $\mathbf{h}$, so $\mathcal{R}_\mathbf{s} = \{\mathbf{h} : \Psi_{\mathrm{LIF}}(\mathbf{h}) = \mathbf{s}\}$ is a polyhedral cell, with strict or nonstrict faces determined by the threshold convention. For stacked patterns, the regions are products of these cells, and the complete spike-path law is $\mathbb{P}[\Psi^{(M)}(\mathbf{Z}) = \mathbf{s}] = \mathbb{P}[\mathbf{Z} \in \mathcal{R}_\mathbf{s}]$.

Boundary-null assumption. At every limiting layer, the Gaussian field assigns probability zero to the union of LIF threshold boundaries. Positive variance of each threshold-defining affine functional suffices, so singular fields, such as rate-projected ones, are admitted unless they are supported on a threshold face.

Under this assumption, convergence of Gaussian means and covariances gives convergence of all cell probabilities. With finitely many patterns, the path law converges in total variation and $\mathcal{T}_\ell$ is continuous. For concentration, conditional iid sampling and bounded spikes give

$$\mathbb{P}\big(\|c_n^{\ell+1} - \mathcal{T}_\ell(c_n^\ell)\|_\infty > \varepsilon \mid \mathcal{F}_\ell\big) \le 2(D + D^2)e^{-2n_{\ell+1}\varepsilon^2}. \tag{4}$$

This controls sampling error about the conditional update, not a global $n^{-1/2}$ bound after many discontinuous layers.

Theorem 3.2 (Wide hard-LIF recursion)**.** *Fix M, T, L. Suppose $c_n^0 \xrightarrow{p} c^0$ deterministically, each output width tends to infinity, and the boundary-null assumption holds along the limiting recursion. Then, at each hidden layer,*

$$c_n^\ell \xrightarrow{p} c^\ell, \qquad c^{\ell+1} = \mathcal{T}_\ell(c^\ell). \tag{5}$$

Any fixed number of neurons converge jointly to independent copies of the path law $\mathcal{L}[\Psi^{(M)}(\mathbf{Z}_\ell(c^{\ell-1}))]$.

The proof combines (4) with continuity by finite induction (Appendix A.1); no surrogate or smoothing of the hard threshold enters.

An independent Gaussian linear readout $f_n(x) = \sigma_{\mathrm{out}} n_L^{-1/2} \sum_i g_i \mathbf{a}^\top \mathbf{S}_i^L(x)$, with iid $g_i \sim \mathcal{N}(0,1)$ and fixed $\mathbf{a} \in \mathbb{R}^T$, converges on every finite input set to a centered Gaussian law with kernel $K_{\mathrm{SNN}}(x, x') = \sigma_{\mathrm{out}}^2 \mathbf{a}^\top \mathbf{K}^L(x, x')\mathbf{a}$. This is the induced NNGP; it is estimated by Gaussian sampling and LIF simulation, not by enumerating the 2^{MT} cells. A rate readout takes $\mathbf{a} = T^{-1}\mathbf{1}_T$, but its kernel can still depend on temporal information processed by earlier layers. Whether it does is a property of the recursion (5), not of the readout alone.

3.4 When the network computes with rates

We compare $\mathcal{T}_\ell$ with the rate-projected map $\mathcal{T}_{\ell,\mathrm{dc}}$ obtained by replacing $\mathbf{B}_\ell$ with $\mathbf{B}_\ell\mathbf{\Pi}$, which sees only the rate quotient. The simplest case is that in which the synapse itself removes every nonconstant direction.

Theorem 3.3 (Exact rate factorization). *If $\mathbf{B}_\ell \mathbf{\Pi}_\perp = \mathbf{0}$ (equivalently $\mathbf{A}_\ell \mathbf{P}_\perp = \mathbf{0}$), the next-layer Gaussian law and moment map factor through $\mathcal{R}(c)$. If this holds at every layer, the downstream recursion depends on the initial descriptor only through its rate quotient.*

Indeed, $\mathbf{Bm} = \mathbf{B\Pi m}$ and $\mathbf{BQB}^\top = \mathbf{B\Pi Q\Pi B}^\top$, so (3) depends on c only through $\mathcal{R}(c)$ (Appendix A.2.2). Output paths need not be temporally constant: LIF generates temporal structure, but the next such synapse again reads only the quotient.

Approximate collapse. For general $\mathbf{B}$, couple the full and rate-projected drives with the same $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I}_D)$: $\mathbf{Z} = \mathbf{b} + \mu\mathbf{Bm} + \sigma\mathbf{BQ}^{1/2}\boldsymbol{\xi}$ and $\mathbf{Z}_{\mathrm{dc}} = \mathbf{b} + \mu\mathbf{B\Pi m} + \sigma\mathbf{B\Pi Q}^{1/2}\boldsymbol{\xi}$. The transmitted temporal leakage energy is exactly $\mathcal{E}_{\mathrm{temp}}(c; \mathbf{B}) = \mathbb{E}\|\mathbf{Z} - \mathbf{Z}_{\mathrm{dc}}\|_2^2 = \mu^2\|\mathbf{B\Pi}_\perp \mathbf{m}\|_2^2 + \sigma^2 \operatorname{tr}(\mathbf{B\Pi}_\perp \mathbf{Q\Pi}_\perp \mathbf{B}^\top)$. It couples the temporal content of the population to the filter that transmits it: large temporal mass near the null space of $\mathbf{B}$ leaks little. Because the threshold is discontinuous, small leakage translates into a small moment error only when the rate-projected field rarely sits near a threshold boundary.

Theorem 3.4 (Approximate hard-LIF collapse). *Let $\mathfrak{B}$ be the LIF boundary union and suppose $\mathbb{P}[\operatorname{dist}(\mathbf{Z}_{\mathrm{dc}}, \mathfrak{B}) \le r] \le C_{\mathrm{ac}} r$ for $0 < r \le r_0$. Then*

$$\|\mathcal{T}(c) - \mathcal{T}_{\mathrm{dc}}(c)\|_\infty \le \inf_{0<r\le r_0} \left(C_{\mathrm{ac}} r + \frac{\mathcal{E}_{\mathrm{temp}}}{r^2} \right). \tag{6}$$

When $(2\mathcal{E}_{\mathrm{temp}}/C_{\mathrm{ac}})^{1/3} \le r_0$, the bound is $3 \cdot 2^{-2/3} C_{\mathrm{ac}}^{2/3} \mathcal{E}_{\mathrm{temp}}^{1/3}$.

A spike path changes only when the coupled fields straddle a boundary; splitting on whether $\mathbf{Z}_{\mathrm{dc}}$ lies within r of it gives (6), which also holds for singular fields.

Since $\|\mathbf{m}\|^2 \le D$ and $\operatorname{tr}\mathbf{Q} \le D$, the condition $\|\mathbf{B\Pi}_\perp\|_{\mathrm{op}} \le \varepsilon$ gives $\mathcal{E}_{\mathrm{temp}} \le D(\mu^2 + \sigma^2)\varepsilon^2$ and one-layer error $O(\varepsilon^{2/3})$. For finite depth, write $\bar{c}^{\ell+1} = \mathcal{T}_{\ell,\mathrm{dc}}(\bar{c}^\ell)$, $\bar{c}^0 = c^0$, and $e_\ell = \|c^\ell - \bar{c}^\ell\|_\infty$. On compact admissible descriptor sets containing both trajectories, assume the projected maps have a common modulus $\omega_\ell(r) \to 0$ and uniform anti-concentration constants $(C_{\mathrm{ac},\ell}, r_{0,\ell})$. Then, whenever $(2\mathcal{E}_{\mathrm{temp}}(c^\ell; \mathbf{B}_\ell)/C_{\mathrm{ac},\ell})^{1/3} \le r_{0,\ell}$, $e_{\ell+1} \le \eta_\ell + \omega_\ell(e_\ell)$, where $\eta_\ell = 3 \cdot 2^{-2/3} C_{\mathrm{ac},\ell}^{2/3} \mathcal{E}_{\mathrm{temp}}(c^\ell; \mathbf{B}_\ell)^{1/3}$. Uniformly vanishing leakage thus gives a vanishing readout-kernel gap at fixed depth (Appendix A.2.8); pointwise boundary-nullness along one trajectory does not supply the uniform modulus, and the limit is in leakage, not depth. Replacing $\mathbf{\Pi}$ by block averages over windows of Δ bins gives a family of coarser projections; the largest Δ^* whose projected kernel stays within a tolerance of K_{SNN} serves as the network's temporal resolution (Appendix A.2.12).

Collapse inside the neuron. The results above constrain the synapse; ANN-to-SNN conversion instead obtains rate sufficiency from the neuron. At the integrator endpoint $\lambda = 1$, $\rho = \vartheta$, summing (2) gives $\vartheta N_T = \sum_{t \le T} h_t + u_0 - u_T$, with N_T the spikes fed back by time T, so the rate is a clipped function of the summed drive up to $|u_T - u_0|/(\vartheta T)$ and the rate quotient closes at $O(1/T)$ when this residual stays bounded (Proposition A.17). Appendix A.3.2 separates this quotient closure from Theorem 3.3 and from equal count laws, and shows that late equal-count spikes can break the residual bound.

3.5 When timing is irreducible: an equal-marginal witness

To show that the collapse conditions cannot be dropped, we need rate-equivalent inputs that the network nevertheless separates. Rate equivalence means equal complete spike-count laws; strong rate equivalence also requires equal per-bin marginals, so a difference cannot come from time-dependent rates and isolates temporal dependence.

For a centered Gaussian pair with positive marginal variances and correlation r, coordinatewise zero-threshold spiking gives

$$\mathbb{E}[\mathbf{1}\{Z_t \ge 0\}\mathbf{1}\{Z_q \ge 0\}] = \frac{1}{4} + \frac{1}{2\pi} \arcsin r. \tag{7}$$

Thus differing normalized correlations of $\sigma^2\mathbf{BQB}^\top$ are sufficient to separate output temporal kernels at the memoryless hard-spiking point $\lambda = \rho = \vartheta = 0$, with zero drive mean. A filtered covariance difference is not enough: it must alter a correlation visible to the threshold, and the operator-norm gap is then at least the corresponding entry difference in (7).

Theorem 3.5 (Strongly rate-equivalent separation). *Let $T = 4$. Let P be uniform on the six two-spike paths and $\widetilde{P}$ uniform on the four paths supported on $\{12, 23, 34, 14\}$. Both have exactly two spikes per sample and $m_t = 1/2$ at every time. Under $\mathbf{B} = \mathbf{I}$, $\mu = 0$, $\mathbf{b} = \mathbf{0}$, $\sigma > 0$, and $\lambda = \rho = \vartheta = 0$, their output temporal kernels differ, and*

$$\mathbf{v}_0^\top (\widetilde{\mathbf{Q}}_{\mathrm{out}} - \mathbf{Q}_{\mathrm{out}})\mathbf{v}_0 = \frac{8\arcsin(1/2) - 12\arcsin(1/3)}{8\pi} \approx 4.406 \times 10^{-3} > 0. \tag{8}$$

The actual firing-rate second-moment gap is one quarter of this value.

For P, every off-diagonal drive correlation is $1/3$. For $\widetilde{P}$, eight ordered off-diagonal pairs have correlation $1/2$ and four have correlation zero. The diagonal output moments agree at $1/2$. Equation (8) follows by summing (7); strict convexity of $\arcsin$ on these nonnegative correlations gives its sign. The cycle covariance is singular, but every coordinate has variance $\sigma^2/2$, so the boundary-null assumption holds. At $T = 2$ the count law and marginals determine the joint law, so strong witnesses need more bins; a two-bin construction and a general witness criterion are in Appendix A.3. This is nonlinear mode conversion: equal input count laws and per-bin rates become different output count second moments, so structure the rate quotient discards at one layer reappears in it at the next. An output rate readout therefore need not make the upstream computation rate sufficient. By continuity the gap persists on an open neighborhood of this point, and it stays positive at all 224 tested LIF settings (Section 4).

From kernel separation to the readout. The collapse and witness results are complementary sufficient conditions, not an if-and-only-if classification: an architecture can transmit temporal energy to which its thresholds or readout are insensitive. A temporal-kernel difference must survive the readout contraction, as (8) verifies for the rate readout, and it affects a task only when the readout selects the surviving temporal directions and the target loads on them. A positive-semidefinite readout weighting $\mathbf{\Lambda}$ defines $K_{\mathbf{\Lambda}}(x, x') = \mathrm{tr}(\mathbf{\Lambda}\mathbf{K}(x, x'))$ and its observable function space $V_{\mathbf{\Lambda}}$; the deficiency theorem lower-bounds prediction risk by the component of the target outside $V_{\mathbf{\Lambda}}$. For an exact rate-null encoder, the projected-network space $V_{\text{rate-proj}}$ contains only constants, so no predictor using that surrogate kernel can beat the best constant while the full temporal kernel separates the inputs (Corollary A.30). This is a conditional observability statement, not an unconditional ordering of finite-sample risks, which also depend on noise and regularization (Appendix A.3.10). Together, these steps organize the tests that follow. Along $\mathbf{Q}_{\mathrm{temp}} \xrightarrow{\mathbf{B}} \mathbf{\Sigma}_{\mathrm{temp}} \xrightarrow{\Psi_{\mathrm{LIF}}} \mathbf{Q}_{\mathrm{out}} \xrightarrow{\text{readout}} K_{\mathrm{SNN}}$, factoring through the rate quotient suffices for rate computation, while a nonconstant component that survives every stage and changes the readout kernel witnesses timing. Neither condition is necessary; together they bracket the two regimes. The operational distinction is therefore survival *and* observability of temporal information.

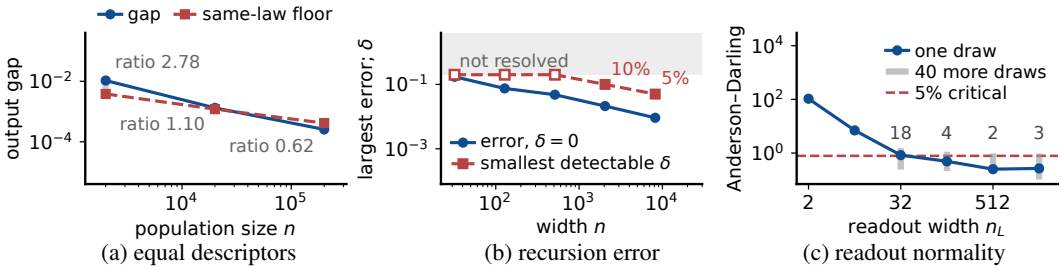
4 Experiments

We check predictions P1–P6 of Appendix A.4.12, which follow the steps of Section 3, and use its quantities as diagnostics on converted and trained checkpoints, where the random-weight theorems need not hold, with the exact hard-LIF map and no forward-pass surrogate outside Appendix A.4.2. Surrogates share draws or weights with their networks, and recalibrated arms share a firing rate. Appendix Table 3 gives the decisive test and result for each link in the argument; Appendix A.4 maps each panel to its protocol.

The limit. Finite hard-LIF networks realise the limit, and tested recursions contract locally (P1). The recursion error falls from 0.173 at width 32 to 0.009 at 8192, equal descriptors stay within 0.62–2.78 times the same-law floor, and the readout approaches its Gaussian limit (Figure 2). All 280 tested recursions contract locally (Figure 3a). We next ask whether the rate-projected surrogate reproduces the limit (P2).

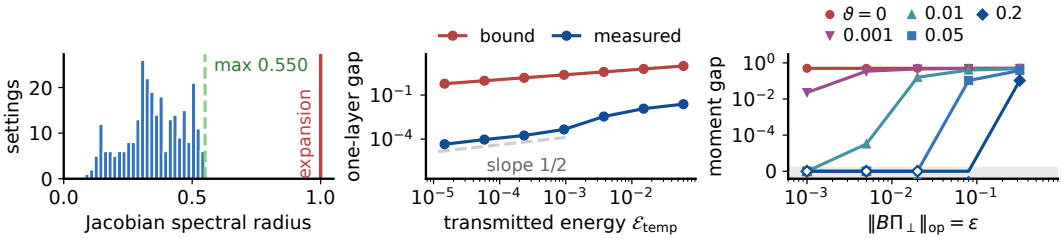
Rate collapse. The surrogate obeys Theorem 3.4 down to a filter-set resolution and fails only at the tangent threshold (P2). The bound lies 397 to 1.3×10^4 times above the gap (Figure 3b), so it certifies consistency, not the gap's size; the gap stays at $1/2$ only at $\vartheta = 0$ (c), Δ^* grows with τ_{syn} (Figure 4a), and T times the integrator's rate-readout gap stays below 0.14 (b). We next ask whether what survives separates inputs that rates cannot (P3–P5).

Separation and transmission. Surviving structure separates rate-equivalent inputs, and transmission and task alignment predict its effect (P3–P5). The equal-marginal pair stays separated at all



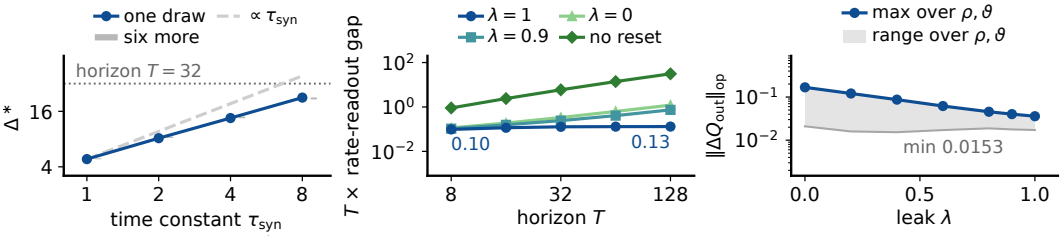
(a) equal descriptors (b) recursion error (c) readout normality

Figure 2: **Finite-width checks of the random hard-LIF limit.**



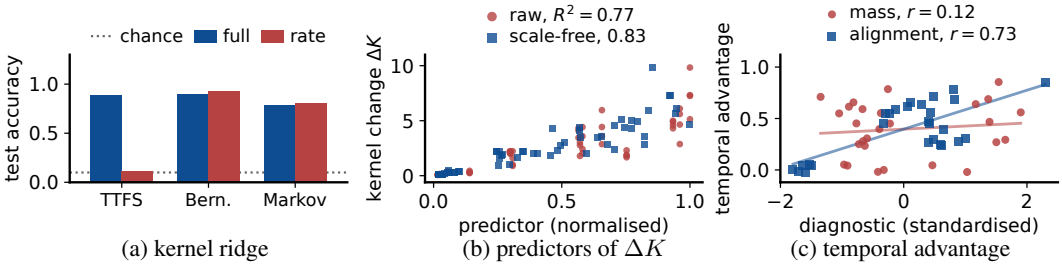
(a) spectral radii (b) one-layer gap, bound (c) gap vs. ε

Figure 3: **Depth contraction, the collapse bound, and its tangent exception.**



(a) resolution Δ^* (b) $T\times$ readout gap (c) equal-marginal pair

Figure 4: **Resolution, integrator scaling, and equal-marginal separation.**



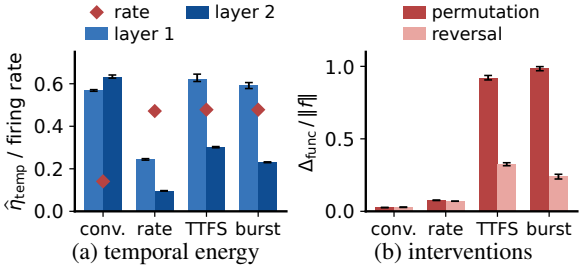
(a) kernel ridge (b) predictors of ΔK (c) temporal advantage

Figure 5: **Functional consequences of transmitted temporal structure.**

224 settings (Figure 4c). Under the exact time-to-first-spike (TTFS) rate null, the rate kernel is constant and sits at chance, 0.110 against 0.890 for the temporal kernel (Figure 5a), instantiating Corollary A.30; the transmitted difference explains the kernel change at $R^2 = 0.83$ against 0.77 (b), and alignment tracks the advantage of 27 trained networks at $r = 0.73$ against 0.12 for temporal mass (c). We next ask whether trained networks fall into the two regimes (P6).

Converted versus native networks. Timing-trained networks depend on timing and rate-derived ones do not, but transmitted energy does not show it (P6). A rate-preserving permutation sends TTFS- and burst-trained networks to chance, changing their logits by 0.92 and 0.99 of their norm against at most 0.08 for the others (Figure 6b), while the converted network's second-layer energy, 0.63, exceeds the TTFS network's 0.30 (a). In this diagnostic, with $\mu = 0$ and $\mathbf{B} = \mathbf{I}$, energy depends on spike counts alone.

Checkpoints as the theory's networks. Eleven published checkpoints from vision, language, speech and control become the hard-LIF networks of the theory: CIFAR ResNets (He et al., 2016; Chen, 2019), ImageNet-1k ResNet-18 (Deng et al., 2009), OPT models on WikiText-2 (Zhang et al., 2022; Merity et al., 2017), T5-small on WMT14 English–German (Raffel et al., 2020; Bojar et al.,



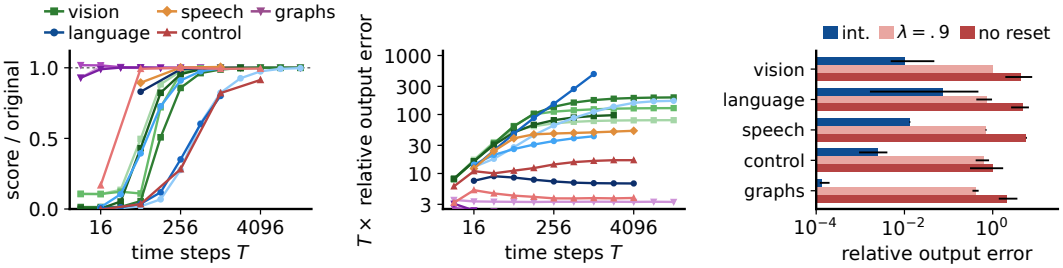
(a) temporal energy (b) interventions

Figure 6: **Computational regimes in trained spiking networks.**

Table 1: **Exact rate-null checkpoints.**

task	checkpoint	permuted	rate surrogate
digits/TTFS	0.9598	0.0787	0.0888
digits/burst	0.9715	0.0938	0.1055
MNIST/TTFS	0.9005	0.1125	0.1170
F-MNIST/TTFS	0.8260	0.0955	0.0970
CIFAR-10/TTFS	0.3500	0.0950	0.0925
text order	0.6515	0.4920	0.5000
SHD spikes	0.737	0.271	0.257

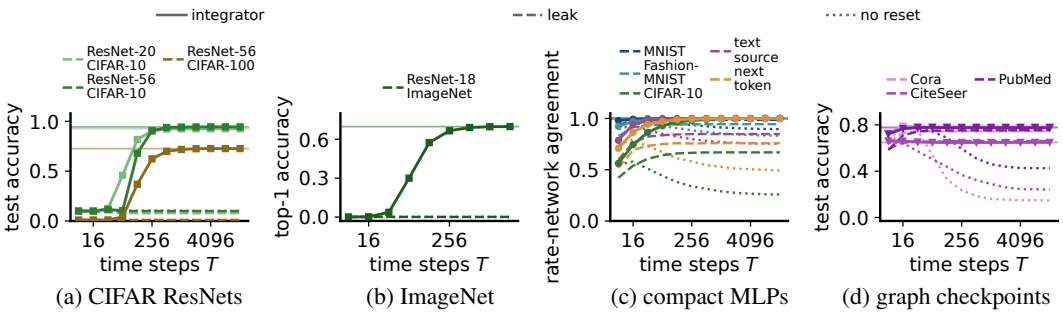
2014), S2T-small on LibriSpeech (Wang et al., 2020; Panayotov et al., 2015), and DQN agents for Pong and Breakout (Mnih et al., 2015; Bellemare et al., 2013; Raffin et al., 2021). For the graph experiments we train three matched GCN checkpoints (Kipf & Welling, 2017) on the standard Planetoid splits of Cora, CiteSeer and PubMed (Yang et al., 2016), rather than relying on one isolated Cora run. Every activation that enters a weight matrix becomes a population of neurons (2) with $u_0 = 0$, two for a signed activation, each unit's threshold is its largest activation on training data folded into the weights, and normalisation and attention act on accumulated spike counts, so the rate network is the checkpoint with every unit clipped at its threshold and reproduces its scores. Every arm runs on the full standard test set: the integrator $(\lambda, \rho) = (1, \vartheta)$ of Proposition A.17 and, as controls, leak $\lambda = 0.9$ and no reset $\rho = 0$ (Appendix A.4.9).



(a) task recovery (b) $T\times$ output error (c) failed controls

Figure 7: **Converted-checkpoint convergence and failed-hypothesis controls.**

Integrator endpoint. Converted checkpoints approach their original scores as T grows, and their output error falls as $1/T$ once spikes have crossed the network (Figure 7a, b). ImageNet ResNet-18 reaches top-1 0.698 against 0.698 at $T = 2048$ (Figure 8b), T5 BLEU 24.1 against 24.0 and S2T comes within 0.6 points of the original word error rate at $T = 1024$ (Table 2b), and the Breakout agent scores 345 against 377 points at $T = 4096$ (c). At $T = 16384$ the three GCNs recover test accuracies 0.783, 0.649 and 0.776 and agree with their rate networks on every test node (Figure 8d). Language models need the longest horizons: at $T = 1024$ OPT-350M is 0.5% above its perplexity while OPT-125M and OPT-1.3B are 22 and 25% above theirs, and at $T = 16384$ OPT-125M is 0.24% above on the whole test set and OPT-1.3B 0.22% above on its first window, both with T times the error levelled off (Figure 7b); this is asymptotic equivalence, not low latency.



(a) CIFAR ResNets (b) ImageNet (c) compact MLPs (d) graph checkpoints

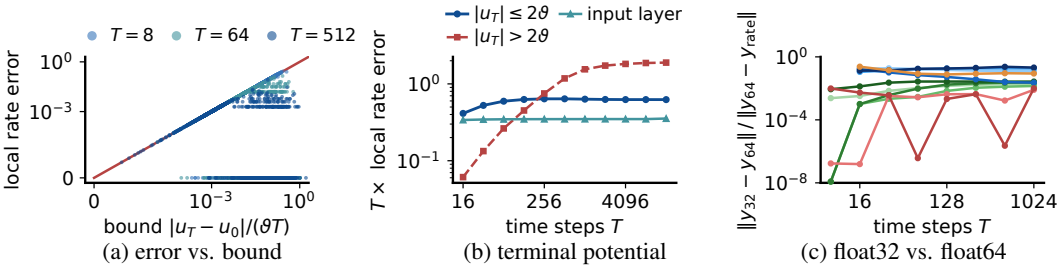
Figure 8: **Vision, compact-MLP and graph-checkpoint task trajectories.**

Controls. Without the proposition's hypotheses the output error settles at a nonzero value instead of falling. At the largest common T it stays between 0.35 and 1.00 with leak and between 0.31 and 7.9 without reset (Figure 7c), which sends the ResNets to chance, both agents to the return of

Table 2: **Sequence and control checkpoint task values.**

(a) OPT, WikiText-2 perplexity↓			(b) T5 BLEU↑, S2T WER↓			(c) Atari return↑		
model @ T	checkpoint rate	SNN	task @ T	checkpoint rate	SNN	agent @ T	checkpoint rate	SNN
	27.656			24.04			20.6±0.5	
125M @ 1k	27.655	33.61	T5 @ 64	24.07	19.96	Pong @ 16	20.6	−13.8
	27.656			24.04			20.6±0.5	
125M @ 16k	27.655	27.723	T5 @ 1k	24.07	24.09	Pong @ 4k	20.6	20.4±0.6
	22.002			5.12			377.1±34.0	
350M @ 1k	22.002	22.11	S2T @ 64	5.09	15.17	Breakout @ 16	376.6	0.2
	14.624			5.12			377.1±34.0	
1.3B @ 1k	14.624	18.26	S2T @ 1k	5.09	4.56	Breakout @ 4k	376.6	344.8±80.9

a random policy, the language models far above their perplexities and the decoded translations and transcriptions to BLEU 0.1 and word error rate 100% (Figure 8 and Appendix Figure 18). Across the three graph datasets, leak leaves relative output error 0.35–0.48 and only 0.846–0.908 agreement with the rate network, while no reset leaves error 1.37–3.62 and agreement 0.153–0.425 (Figure 8d). The compact MLPs show the same separation: agreement at the largest plotted T is 0.669–0.986 with leak and 0.258–0.896 without reset, versus at least 0.9995 for every integrator (Figure 8c).



(a) error vs. bound (b) terminal potential (c) float32 vs. float64

Figure 9: **Exact mechanism and numerical checks for the integrator endpoint.**

Exact checks. The identity and the bound of Proposition A.17 hold on every neuron checked, with no violation in 2.2×10^{10} published-checkpoint pairs or any GCN check; maximum residuals are 1.2×10^{-10} and 7.3×10^{-9}, respectively (Figure 9a). Single precision changes the published-checkpoint output by an amount that falls with T like the error itself and stays below a quarter of it (c). Neurons whose potential stays within 2ϑ keep T times their local error bounded, whereas sign-changing drives or late accumulated input can make the terminal residual grow; these neurons carry the transient (b).

Exact rate null. Timing-code checkpoints are irreducible exactly as the rate quotient predicts. Their same-weight rate surrogates return bitwise identical logits on rate-equivalent inputs, so accuracy equals its predicted value exactly, 0.0888 on digits and 0.5000 on word order, against 0.35–0.97 and 0.65 for the checkpoints, and on the Spiking Heidelberg Digits it is invariant to permuting each channel's spike times, at 0.26 against 0.74 (Table 1). In these networks, timing matters when its second-order structure survives the architecture and reaches a readout that uses it.

5 Conclusion

Gaussian synaptic mixing makes $(\mathbf{m}, \mathbf{Q})$ an exact sufficient description of the next-layer temporal channel. At fixed depth and horizon, hard-LIF boundary geometry and concentration turn that channel into a deterministic temporal spike-kernel recursion and a Gaussian-process readout. Rate coding retains its constant temporal quotient. Suppressing transmitted complementary directions guarantees rate collapse; a strict equal-marginal construction shows that surviving temporal dependence can change even a downstream rate kernel. Numerically, finite networks support the recursion, collapse and separation predictions, integrator endpoint, and exact rate null. The claims cover random Gaussian networks, boundary-null fields, and fixed finite horizons; non-Gaussian weights, growing horizons, recurrent synaptic reuse, and trained-network theory remain open. For trained networks these are diagnostics: timing matters there only through statistics the architecture preserves and its output sees.

AI USE STATEMENT

In this work, we used generative AI tools to polish the text, to search for related work, and to create the figures. All figures were drawn by coding agents (Claude Code and Codex) following figure-making skills, i.e., instruction files that fix the house style, the panel layout and automatic legibility checks. Each data figure is produced by a matplotlib script that reads the archived experimental outputs, and the schematic in Figure 1 is an editable vector diagram; no image-generation model was used. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58, 2014.

Romain Brette. Philosophy of the spike: rate-based vs. spike-based theories of the brain. *Frontiers in Systems Neuroscience*, 9:151, 2015.

Tong Bu, Wei Fang, Jianhao Ding, Penglin Dai, Zhaofei Yu, and Tiejun Huang. Optimal ANN-SNN conversion for high-accuracy and ultra-low-latency spiking neural networks. In *International Conference on Learning Representations*, 2022.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):2914, 2021.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Yaofo Chen. PyTorch CIFAR models. `https://github.com/chenyaofo/pytorch-cifar-models`, 2019. Accessed 2026-09-13.

Benjamin Cramer, Yannik Stradmann, Johannes Schemmel, and Friedemann Zenke. The Heidelberg spiking data sets for the systematic evaluation of spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2744–2757, 2022.

Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz Kandola. On kernel-target alignment. In *Advances in Neural Information Processing Systems*, 2001.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018.

Wolfgang Maass. Lower bounds for the computational power of networks of spiking neurons. *Neural Computation*, 8(1):1–40, 1996.

Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural Networks*, 10(9):1659–1671, 1997.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Ernest Montbrió, Diego Pazó, and Alex Roxin. Macroscopic description for networks of spiking neurons. *Physical Review X*, 5(2):021028, 2015.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

Stefano Panzeri, Christopher D. Harvey, Eugenio Piasini, Peter E. Latham, and Tommaso Fellin. Cracking the neural code for sensory perception by combining statistics, intervention, and behavior. *Neuron*, 93(3):491–507, 2017.

Stefano Panzeri, Nicola Marie Engel, and Marco Celotto. Contribution of spike timing to the neural code: from fast to slow timescales. *Biological Cybernetics*, 120(2):12, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems*, 2016.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, 2018.

Robert Price. A useful theorem for nonlinear devices having Gaussian inputs. *IRE Transactions on Information Theory*, 4(2):69–72, 1958.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

Bodo Rueckauer, Iulia-Alexandra Lungu, Yuhuang Hu, Michael Pfeiffer, and Shih-Chii Liu. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in Neuroscience*, 11:682, 2017.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations*, 2017.

Ana Stanojevic, Stanisław Woźniak, Guillaume Bellec, Giovanni Cherubini, Angeliki Pantazi, and Wulfram Gerstner. An exact mapping from ReLU networks to spiking neural networks. *Neural Networks*, 168:74–88, 2023.

Frédéric Theunissen and John P. Miller. Temporal encoding in nervous systems: A rigorous definition. *Journal of Computational Neuroscience*, 2(2):149–162, 1995.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 33–39, 2020.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *International Conference on Machine Learning*, pp. 40–48, 2016.

Ziqiao Yu, Pengfei Sun, and Dan F. M. Goodman. Beyond rate coding: Surrogate gradients enable spike timing learning in spiking neural networks. *Neuromorphic Computing and Engineering*, 6 (1):014016, 2026.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A THEORY, EVIDENCE, AND DIAGNOSTICS

This appendix follows a single argument. Section A.1 establishes the temporal representation and its wide-limit recursion; Sections A.2 and A.3 then give complementary sufficient directions, showing when an architecture removes timing information and when rate-equivalent inputs remain distinguishable downstream. Together, these results identify when timing survives the architecture and remains observable at the readout. Section A.4 tests both directions in finite networks and trained checkpoints, while Section A.5 verifies the boundary-nullity and anti-concentration assumptions used by the continuity and collapse arguments. Table 3 maps each empirical claim to its decisive test and finding.

Table 3: **Summary of empirical tests.**

claim	test	finding
P1: limit	Width $32 \to 8192$; depth sweep	Recursion error $0.173 \to 0.009$; $280/280$ settings contract.
P2: collapse	Full versus rate kernel	Bound/gap 397–1.3×10^4; only the tangent case retains gap $1/2$.
P2: resolution	Filter timescale; integrator horizon	Δ^* follows τ_{syn}; T-scaled gap < 0.14.
P3: separation	Fixed counts and marginals	$224/224$ settings remain separated.
P4: transmission	Raw versus transmitted predictor	R^2: $0.77 \to 0.83$.
P5: task use	Exact TTFS rate null	Rate/temporal accuracy $0.110/0.890$.
P5: alignment	Temporal mass versus alignment	Advantage correlation 0.12 versus 0.73.
P6: regimes	Rate-preserving interventions	Relative logit change ≤ 0.08 versus $0.92/0.99$.
P6: energy control	Transmitted energy ordering	Converted $0.63 >$ TTFS 0.30: energy alone fails.
Checkpoint scope	Published models and agents	14 checkpoints: 11 published and 3 matched GCNs.
Integrator scaling	Output error versus T	$O(1/T)$ after the network transient.
Task recovery	Scores at the largest horizon	R18 exact; T5 $24.1/24.0$; S2T within 0.6 WER; Breakout $345/377$.
Failed controls	Leak / no reset	Error floor 0.35–1.00 / 0.31–7.9.
Exact checks	Identity and bound	Published residual 1.2×10^{-10}; GCN residual 7.3×10^{-9}; no violation.
Exact rate null	Equal-rate inputs	Rate logits identical; timing checkpoints retain 0.35–0.97 accuracy.

A.1 GAUSSIAN COARSE-GRAINING AND WIDE-LIMIT PROOFS

We now derive the temporal Gaussian-process representation underlying our analysis. Our goal is not merely to establish that a sufficiently wide SNN admits a Gaussian-process limit. Rather, we identify the *temporal second-order object* through which the entire downstream computation factors, while allowing the neuron dynamics themselves to remain discontinuous and stateful.

With Gaussian random weights, the complete temporal preactivation path of a neuron is conditionally Gaussian *at finite width*. Its conditional mean and covariance depend on the preceding spiking population only through its empirical first and second temporal moments. The infinite-width limit is then needed only to make these empirical quantities self-average to deterministic values. This produces a temporal analogue of the NNGP recursion in which the hidden kernel is matrix-valued over time.

A.1.1 FINITE-INPUT TEMPORAL REPRESENTATION

Fix a finite collection of inputs $\mathcal{X}_M = \{x^1, \dots, x^M\}$ and a finite time horizon T. Throughout the section, $D := MT$. For neuron i at layer ℓ, let $\mathbf{S}_i^\ell(x^a) = \big(S_i^\ell(x^a, 1), \dots, S_i^\ell(x^a, T)\big)^\top \in \{0,1\}^T$ denote its spike trajectory in response to input x^a. We stack the responses to all M inputs into $\mathbf{S}_i^\ell := \mathrm{vec}(\mathbf{S}_i^\ell(x^1), \dots, \mathbf{S}_i^\ell(x^M)) \in \{0,1\}^D$. The empirical first and second temporal moments are $\mathbf{m}_n^\ell := n_\ell^{-1} \sum_{i=1}^{n_\ell} \mathbf{S}_i^\ell \in \mathbb{R}^D$ and $\mathbf{Q}_n^\ell := n_\ell^{-1} \sum_{i=1}^{n_\ell} \mathbf{S}_i^\ell (\mathbf{S}_i^\ell)^\top \in \mathbb{R}^{D \times D}$.

The (a, b)-th $T \times T$ block of $\mathbf{Q}_n^\ell$, $\mathbf{Q}_{n,ab}^\ell$, records the joint temporal activity of a typical neuron under inputs x^a and x^b. In the infinite-width limit we will write $\mathbf{K}^\ell(x^a, x^b) := \mathbf{Q}_{ab}^\ell \in \mathbb{R}^{T\times T}$ and refer to $\mathbf{K}^\ell$ as the *temporal spike kernel*.

Let $\mathbf{A}_\ell \in \mathbb{R}^{T\times T}$ denote the temporal synaptic operator at layer ℓ, and define its blockwise extension $\mathbf{B}_\ell := \mathbf{I}_M \otimes \mathbf{A}_\ell \in \mathbb{R}^{D\times D}$. Thus the filtered trajectory of neuron j is $\mathbf{X}_j^\ell = \mathbf{B}_\ell \mathbf{S}_j^\ell$. We use the mean-plus-fluctuation parameterization

$$W_{ij}^{\ell+1} = \frac{\mu_{\ell+1}}{n_\ell} + \frac{\sigma_{\ell+1}}{\sqrt{n_\ell}} G_{ij}^{\ell+1}, \qquad G_{ij}^{\ell+1} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1). \tag{9}$$

The complete preactivation trajectory of neuron i at layer $\ell + 1$, evaluated jointly on all M inputs, is therefore

$$\mathbf{H}_i^{\ell+1} = \mathbf{b}_{\ell+1} + \sum_{j=1}^{n_\ell} W_{ij}^{\ell+1} \mathbf{B}_\ell \mathbf{S}_j^\ell, \tag{10}$$

where $\mathbf{b}_{\ell+1} \in \mathbb{R}^D$ is a deterministic bias vector.

A.1.2 Exact conditional Gaussianity

The first result is finite-width and exact.

Proposition A.1 (Exact conditional Gaussianity)**.** *Let $\mathcal{F}_\ell := \sigma\left(\mathbf{S}_1^\ell, \ldots, \mathbf{S}_{n_\ell}^\ell\right)$ be the sigma-field generated by the spike trajectories at layer ℓ. Under the weight parameterization* (9)*,*

$$\mathbf{H}_i^{\ell+1} \mid \mathcal{F}_\ell \sim \mathcal{N}\left(\mathbf{M}_{\ell+1}^{(n)}, \mathbf{\Sigma}_{\ell+1}^{(n)}\right), \tag{11}$$

where $\mathbf{M}_{\ell+1}^{(n)} = \mathbf{b}_{\ell+1} + \mu_{\ell+1} \mathbf{B}_\ell \mathbf{m}_n^\ell$, and

$$\mathbf{\Sigma}_{\ell+1}^{(n)} = \sigma_{\ell+1}^2 \mathbf{B}_\ell \mathbf{Q}_n^\ell \mathbf{B}_\ell^\top. \tag{12}$$

Moreover, conditioned on $\mathcal{F}_\ell$, the vectors $\{\mathbf{H}_i^{\ell+1}\}_{i=1}^{n_{\ell+1}}$ are independent and identically distributed.

Proof. Substituting (9) into (10) gives

$$\mathbf{H}_i^{\ell+1} = \mathbf{b}_{\ell+1} + \mu_{\ell+1} \mathbf{B}_\ell \left(\frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \mathbf{S}_j^\ell\right) + \frac{\sigma_{\ell+1}}{\sqrt{n_\ell}} \sum_{j=1}^{n_\ell} G_{ij}^{\ell+1} \mathbf{B}_\ell \mathbf{S}_j^\ell.$$

Conditioned on $\mathcal{F}_\ell$, the vectors $\mathbf{B}_\ell \mathbf{S}_j^\ell$ are deterministic. The final term is therefore a linear combination of independent standard Gaussian variables and is exactly Gaussian.

Its conditional expectation is zero and its conditional covariance is $\mathrm{Cov}[\sigma_{\ell+1} n_\ell^{-1/2} \sum_j G_{ij}^{\ell+1} \mathbf{B}_\ell \mathbf{S}_j^\ell \mid \mathcal{F}_\ell] = \sigma_{\ell+1}^2 n_\ell^{-1} \sum_{j=1}^{n_\ell} \mathbf{B}_\ell \mathbf{S}_j^\ell (\mathbf{S}_j^\ell)^\top \mathbf{B}_\ell^\top = \sigma_{\ell+1}^2 \mathbf{B}_\ell \mathbf{Q}_n^\ell \mathbf{B}_\ell^\top$.

This proves (11)–(12). Independence across output neurons follows from independence of the Gaussian weight rows $\{\mathbf{G}_{i:}^{\ell+1}\}_i$. □

Proposition A.1 yields an exact finite-width coarse-graining property: $\mathcal{L}(\mathbf{H}_i^{\ell+1} \mid \mathbf{S}_1^\ell, \ldots, \mathbf{S}_{n_\ell}^\ell) = \mathcal{L}(\mathbf{H}_i^{\ell+1} \mid \mathbf{m}_n^\ell, \mathbf{Q}_n^\ell)$. Thus two microscopic spike populations with identical $(\mathbf{m}_n^\ell, \mathbf{Q}_n^\ell)$ are indistinguishable to the next random wide layer, even if their third- and higher-order empirical spike statistics differ.

This second-order sufficiency is not an assumption imposed by our analysis. It follows exactly from Gaussian random mixing.

A.1.3 Hard LIF as a Gaussian polyhedral channel

We next characterize the effect of hard LIF dynamics on a Gaussian temporal field.

For a single input, let $\mathbf{h} = (h_1, \ldots, h_T)^\top \in \mathbb{R}^T$ be the input-drive path. The LIF dynamics are

$$u_t = \lambda u_{t-1} + h_t - \rho s_{t-1}, \qquad s_t = \mathbf{1}\{u_t \geq \vartheta\}, \tag{13}$$

with fixed initial state u_0 and $s_0 = 0$. These equations define the deterministic path map $\Psi_{\mathrm{LIF}} : \mathbb{R}^T \to \{0,1\}^T$. Although Ψ_{LIF} is discontinuous, its discontinuity structure is particularly simple.

Lemma A.2 (Polyhedral spike regions)**.** *For every spike pattern $\mathbf{s} = (s_1, \ldots, s_T) \in \{0,1\}^T$, the set $\mathcal{R}_{\mathbf{s}} := \{\mathbf{h} : \Psi_{\mathrm{LIF}}(\mathbf{h}) = \mathbf{s}\}$ is a polyhedral cell up to the convention used on its boundary. In particular, its boundary is contained in a finite union of affine hyperplanes.*

Proof. Condition on a candidate spike trajectory $\mathbf{s}$. Recursively unrolling (13) gives $u_t = \lambda^t u_0 + \sum_{q=1}^{t} \lambda^{t-q} h_q - \rho \sum_{q=1}^{t-1} \lambda^{t-1-q} s_q$. Define the lower-triangular matrix $\mathbf{L}_\lambda \in \mathbb{R}^{T\times T}$ by $(\mathbf{L}_\lambda)_{tq} = \lambda^{t-q}$ for $q \le t$ and $(\mathbf{L}_\lambda)_{tq} = 0$ for $q > t$, and define

$$c_t(\mathbf{s}) = \vartheta - \lambda^t u_0 + \rho \sum_{q=1}^{t-1} \lambda^{t-1-q} s_q. \tag{14}$$

Then $u_t - \vartheta = (\mathbf{L}_\lambda \mathbf{h})_t - c_t(\mathbf{s})$. For $s_t = 1$, consistency with the candidate spike pattern requires $(\mathbf{L}_\lambda \mathbf{h})_t - c_t(\mathbf{s}) \ge 0$, whereas for $s_t = 0$ it requires $(\mathbf{L}_\lambda \mathbf{h})_t - c_t(\mathbf{s}) < 0$. Hence $\mathcal{R}_{\mathbf{s}}$ is the intersection of finitely many affine half-spaces, with only the strict-versus-nonstrict boundary convention determined by (13). Its boundary is contained in $\bigcup_{t=1}^{T} \{\mathbf{h} : (\mathbf{L}_\lambda \mathbf{h})_t = c_t(\mathbf{s})\}$, which is a finite union of affine hyperplanes. □

For M inputs we apply the same map independently to each input block and write $\Psi_{\mathrm{LIF}}^{(M)} : \mathbb{R}^D \to \{0,1\}^D$. For a stacked spike pattern $\mathbf{s} = (\mathbf{s}^1, \ldots, \mathbf{s}^M)$, the corresponding region is $\mathcal{R}_{\mathbf{s}} = \mathcal{R}_{\mathbf{s}^1} \times \cdots \times \mathcal{R}_{\mathbf{s}^M}$, and is again polyhedral up to boundary conventions.

If $\mathbf{Z} \sim \mathcal{N}(\mathbf{M}, \mathbf{\Sigma})$, then the complete joint distribution of the output spike trajectories is therefore $\mathbb{P}[\Psi_{\mathrm{LIF}}^{(M)}(\mathbf{Z}) = \mathbf{s}] = \mathbb{P}[\mathbf{Z} \in \mathcal{R}_{\mathbf{s}}]$. Thus hard LIF transforms Gaussian temporal fields through Gaussian probabilities over polyhedral cells.

No differentiability of the threshold map is required.

A.1.4 The temporal moment map

Let $c = (\mathbf{m}, \mathbf{Q})$ denote a candidate pair consisting of a first moment $\mathbf{m} \in \mathbb{R}^D$ and a positive semidefinite second-moment matrix $\mathbf{Q} \in \mathbb{R}^{D\times D}$.

At layer $\ell + 1$, define

$$\mathbf{M}_{\ell+1}(c) := \mathbf{b}_{\ell+1} + \mu_{\ell+1} \mathbf{B}_\ell \mathbf{m}, \tag{15}$$

and

$$\mathbf{\Sigma}_{\ell+1}(c) := \sigma_{\ell+1}^2 \mathbf{B}_\ell \mathbf{Q} \mathbf{B}_\ell^\top. \tag{16}$$

Let

$$\mathbf{Z}_{\ell+1}(c) \sim \mathcal{N}\left(\mathbf{M}_{\ell+1}(c), \mathbf{\Sigma}_{\ell+1}(c)\right), \tag{17}$$

and define the induced spike vector $\mathbf{S}_{\ell+1}(c) := \Psi_{\mathrm{LIF}}^{(M)}\left(\mathbf{Z}_{\ell+1}(c)\right)$. We define the layerwise temporal moment map by $\mathcal{T}_\ell(\mathbf{m}, \mathbf{Q}) := (\mathcal{T}_\ell^{(1)}(\mathbf{m}, \mathbf{Q}), \mathcal{T}_\ell^{(2)}(\mathbf{m}, \mathbf{Q}))$ with $\mathcal{T}_\ell^{(1)}(\mathbf{m}, \mathbf{Q}) = \mathbb{E}[\mathbf{S}_{\ell+1}(c)]$ and $\mathcal{T}_\ell^{(2)}(\mathbf{m}, \mathbf{Q}) = \mathbb{E}[\mathbf{S}_{\ell+1}(c)\mathbf{S}_{\ell+1}(c)^\top]$.

Equivalently, using the polyhedral decomposition,

$$\mathcal{T}_\ell^{(1)}(\mathbf{m}, \mathbf{Q}) = \sum_{\mathbf{s}\in\{0,1\}^D} \mathbf{s}\, \mathbb{P}\left[\mathbf{Z}_{\ell+1}(c) \in \mathcal{R}_{\mathbf{s}}\right], \tag{18}$$

and

$$\mathcal{T}_\ell^{(2)}(\mathbf{m}, \mathbf{Q}) = \sum_{\mathbf{s}\in\{0,1\}^D} \mathbf{s}\mathbf{s}^\top\, \mathbb{P}\left[\mathbf{Z}_{\ell+1}(c) \in \mathcal{R}_{\mathbf{s}}\right]. \tag{19}$$

The apparent 2^D representation in (18)–(19) is used only for the theoretical characterization. Practical kernel estimation does not require enumerating all spike patterns.

A.1.5 CONTINUITY DESPITE HARD THRESHOLDS

The discontinuity of the hard threshold does not prevent continuity of the moment map at nondegenerate Gaussian laws. We use the following mild assumption.

Assumption 1 (Boundary-null Gaussian fields). For every layer ℓ and every limiting descriptor $c^\ell = (\mathbf{m}^\ell, \mathbf{Q}^\ell)$, $\mathbb{P}\left[\mathbf{Z}_{\ell+1}(c^\ell) \in \partial\mathcal{R}_\mathbf{s}\right] = 0$ for every $\mathbf{s} \in \{0,1\}^D$.

A sufficient condition is that each threshold-defining affine functional in Lemma A.2 has strictly positive variance under the limiting Gaussian field.

Lemma A.3 (Continuity of the hard-LIF moment map)**.** *Suppose* $\mathbf{m}_k \to \mathbf{m}$, $\mathbf{Q}_k \to \mathbf{Q}$, *and Assumption 1 holds at* $c = (\mathbf{m}, \mathbf{Q})$. *Then*

$$\mathcal{T}_\ell(\mathbf{m}_k, \mathbf{Q}_k) \to \mathcal{T}_\ell(\mathbf{m}, \mathbf{Q}). \tag{20}$$

Proof. By linearity of (15)–(16), $\mathbf{M}_{\ell+1}(\mathbf{m}_k, \mathbf{Q}_k) \to \mathbf{M}_{\ell+1}(\mathbf{m}, \mathbf{Q})$ and $\mathbf{\Sigma}_{\ell+1}(\mathbf{m}_k, \mathbf{Q}_k) \to \mathbf{\Sigma}_{\ell+1}(\mathbf{m}, \mathbf{Q})$. Hence the corresponding Gaussian measures converge weakly: $\mathcal{N}(\mathbf{M}_k, \mathbf{\Sigma}_k) \Rightarrow \mathcal{N}(\mathbf{M}, \mathbf{\Sigma})$. For every spike pattern $\mathbf{s}$, Assumption 1 implies that $\mathcal{R}_\mathbf{s}$ is a continuity set of the limiting Gaussian measure. The Portmanteau theorem therefore gives $\mathbb{P}[\mathbf{Z}_k \in \mathcal{R}_\mathbf{s}] \to \mathbb{P}[\mathbf{Z} \in \mathcal{R}_\mathbf{s}]$.

Because the spike-pattern space $\{0,1\}^D$ is finite, all pattern probabilities converge simultaneously. Substituting these probabilities into (18)–(19) yields (20). □

In fact, Lemma A.3 gives convergence of the complete induced spike-path distribution in total variation, since the output space is finite. This observation is useful conceptually: the forward theory does not require replacing the hard threshold by a smooth surrogate. The only problematic points are Gaussian trajectories lying exactly on threshold boundaries, which have zero probability under Assumption 1.

A.1.6 CONDITIONAL CONCENTRATION OF TEMPORAL MOMENTS

Conditioned on the previous layer, Proposition A.1 and the deterministic LIF map imply $\mathbf{S}_1^{\ell+1}, \ldots, \mathbf{S}_{n_{\ell+1}}^{\ell+1} \mid \mathcal{F}_\ell$ are independent and identically distributed. Their conditional first and second moments are exactly $\mathcal{T}_\ell\left(\mathbf{m}_n^\ell, \mathbf{Q}_n^\ell\right)$. This immediately yields a finite-width concentration result.

Lemma A.4 (Conditional moment concentration)**.** *For every* $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\left[\left\|(\mathbf{m}_n^{\ell+1}, \mathbf{Q}_n^{\ell+1}) - \mathcal{T}_\ell(\mathbf{m}_n^\ell, \mathbf{Q}_n^\ell)\right\|_\infty > \varepsilon \mid \mathcal{F}_\ell\right] \\ \leq 2(D + D^2)\exp\left(-2n_{\ell+1}\varepsilon^2\right).\end{aligned} \tag{21}$$

Consequently, $(\mathbf{m}_n^{\ell+1}, \mathbf{Q}_n^{\ell+1}) - \mathcal{T}_\ell(\mathbf{m}_n^\ell, \mathbf{Q}_n^\ell) = O_p(n_{\ell+1}^{-1/2})$ *coordinate-wise for fixed* $D = MT$.

Proof. For every coordinate $r \in \{1, \ldots, D\}$, $m_n^{\ell+1}(r) = \frac{1}{n_{\ell+1}} \sum_{i=1}^{n_{\ell+1}} S_i^{\ell+1}(r)$ is the average of conditionally iid random variables taking values in $[0,1]$. Hoeffding's inequality therefore gives $\mathbb{P}[|m_n^{\ell+1}(r) - \mathbb{E}[S_i^{\ell+1}(r) \mid \mathcal{F}_\ell]| > \varepsilon \mid \mathcal{F}_\ell] \leq 2e^{-2n_{\ell+1}\varepsilon^2}$.

Similarly, for every pair r, q, $Q_n^{\ell+1}(r,q) = \frac{1}{n_{\ell+1}} \sum_i S_i^{\ell+1}(r) S_i^{\ell+1}(q)$, and each summand also lies in $[0,1]$. Applying Hoeffding again and taking a union bound over the $D + D^2$ coordinates yields (21). □

The decomposition $(\mathbf{m}_n^{\ell+1}, \mathbf{Q}_n^{\ell+1}) = \mathcal{T}_\ell(\mathbf{m}_n^\ell, \mathbf{Q}_n^\ell) + O_p(n_{\ell+1}^{-1/2})$ is the temporal analogue of the layerwise kernel concentration underlying conventional NNGP limits.

A.1.7 FINITE-DEPTH TEMPORAL NNGP LIMIT

We can now state the main result of this section.

Theorem A.5 (Temporal NNGP limit for hard spiking networks)**.** *Fix a finite number of inputs* M, *a finite temporal horizon* T, *and a finite network depth* L. *Consider the feedforward SNN defined by* (9)–(13).

Assume that

$$(\mathbf{m}_n^0, \mathbf{Q}_n^0) \xrightarrow{p} (\mathbf{m}^0, \mathbf{Q}^0), \tag{22}$$

and that Assumption 1 holds at every limiting layer.

Then, as $\min_{\ell=1,\dots,L} n_\ell \to \infty$*, for every finite layer* ℓ*,*

$$(\mathbf{m}_n^\ell, \mathbf{Q}_n^\ell) \xrightarrow{p} (\mathbf{m}^\ell, \mathbf{Q}^\ell), \tag{23}$$

where the deterministic limit is recursively defined by

$$(\mathbf{m}^{\ell+1}, \mathbf{Q}^{\ell+1}) = \mathcal{T}_\ell(\mathbf{m}^\ell, \mathbf{Q}^\ell). \tag{24}$$

Furthermore, for every fixed k*,*

$$(\mathbf{S}_1^\ell, \dots, \mathbf{S}_k^\ell) \Rightarrow \Lambda_\ell^{\otimes k}, \tag{25}$$

where

$$\Lambda_\ell = \mathcal{L}\left[\Psi_{\mathrm{LIF}}^{(M)}(\mathbf{Z}_\ell)\right], \tag{26}$$

and $\mathbf{Z}_\ell \sim \mathcal{N}(\mathbf{M}_\ell(\mathbf{m}^{\ell-1}, \mathbf{Q}^{\ell-1}), \mathbf{\Sigma}_\ell(\mathbf{m}^{\ell-1}, \mathbf{Q}^{\ell-1}))$*. Hence neurons at every finite layer become asymptotically independent copies of a deterministic spike-path law whose complete dependence on the preceding layer is mediated by its first and second temporal moments.*

Proof. We proceed by induction over layers.

For the base layer, (22) holds by assumption.

Suppose for some ℓ that $(\mathbf{m}_n^\ell, \mathbf{Q}_n^\ell) \xrightarrow{p} (\mathbf{m}^\ell, \mathbf{Q}^\ell)$. By Lemma A.4,

$$(\mathbf{m}_n^{\ell+1}, \mathbf{Q}_n^{\ell+1}) - \mathcal{T}_\ell(\mathbf{m}_n^\ell, \mathbf{Q}_n^\ell) \xrightarrow{p} 0. \tag{27}$$

By Lemma A.3 and the continuous mapping theorem,

$$\mathcal{T}_\ell(\mathbf{m}_n^\ell, \mathbf{Q}_n^\ell) \xrightarrow{p} \mathcal{T}_\ell(\mathbf{m}^\ell, \mathbf{Q}^\ell). \tag{28}$$

Combining (27) and (28) gives $(\mathbf{m}_n^{\ell+1}, \mathbf{Q}_n^{\ell+1}) \xrightarrow{p} \mathcal{T}_\ell(\mathbf{m}^\ell, \mathbf{Q}^\ell)$, which establishes (23)–(24) by finite induction. It remains to prove (25). Conditioned on $\mathcal{F}_{\ell-1}$, the spike trajectories $\mathbf{S}_1^\ell, \dots, \mathbf{S}_k^\ell$ are iid with common law

$\Lambda_\ell\left(\mathbf{m}_n^{\ell-1}, \mathbf{Q}_n^{\ell-1}\right)$, where $\Lambda_\ell(\mathbf{m}, \mathbf{Q}) := \mathcal{L}[\Psi_{\mathrm{LIF}}^{(M)}(\mathbf{Z}_\ell(\mathbf{m}, \mathbf{Q}))]$ with $\mathbf{Z}_\ell(\mathbf{m}, \mathbf{Q})$ given by (17) at layer ℓ, so that the law in (26) is $\Lambda_\ell = \Lambda_\ell(\mathbf{m}^{\ell-1}, \mathbf{Q}^{\ell-1})$.

Let $f : (\{0,1\}^D)^k \to \mathbb{R}$ be arbitrary. Since the state space is finite, f is bounded. Then $\mathbb{E}[f(\mathbf{S}_1^\ell, \dots, \mathbf{S}_k^\ell)] = \mathbb{E}[F_f(\mathbf{m}_n^{\ell-1}, \mathbf{Q}_n^{\ell-1})]$, where $F_f(\mathbf{m}, \mathbf{Q}) = \int f(\mathbf{s}_1, \dots, \mathbf{s}_k) \prod_{j=1}^k \Lambda_\ell(\mathbf{m}, \mathbf{Q})(d\mathbf{s}_j)$.

By Lemma A.3, F_f is continuous at the limiting descriptor. Since $(\mathbf{m}_n^{\ell-1}, \mathbf{Q}_n^{\ell-1}) \xrightarrow{p} (\mathbf{m}^{\ell-1}, \mathbf{Q}^{\ell-1})$, we obtain $\mathbb{E}[f(\mathbf{S}_1^\ell, \dots, \mathbf{S}_k^\ell)] \to \int f \, d\Lambda_\ell^{\otimes k}$.

Because the state space is finite, this proves convergence in distribution to $\Lambda_\ell^{\otimes k}$. □

Theorem A.5 provides a particularly direct form of layerwise propagation of chaos. Unlike recurrent interacting-particle systems, where propagation of chaos requires a separate coupling argument, here it follows from the feedforward architecture: conditioned on the previous layer, different output neurons already use independent random weight rows. The wide-layer limit is required only to remove randomness from the empirical temporal kernel passed between layers.

A.1.8 Operator-valued temporal kernel recursion

Let $\mathbf{Q}^\ell = \left[\mathbf{Q}_{ab}^\ell\right]_{a,b=1}^M$, where each block $\mathbf{Q}_{ab}^\ell = \mathbf{K}^\ell(x^a, x^b) \in \mathbb{R}^{T\times T}$. Theorem A.5 implies a deterministic recursion $\mathbf{K}^{\ell+1} = \mathcal{K}_\ell(\mathbf{m}^\ell, \mathbf{K}^\ell)$. For the hard-LIF neuron, this kernel map admits an explicit Gaussian-polyhedral representation.

Let $(\mathbf{H}_x, \mathbf{H}_{x'})$ be jointly Gaussian with means $\mathbf{M}_x, \mathbf{M}_{x'}$ and covariance blocks $\mathbf{\Sigma}_{xx}, \mathbf{\Sigma}_{xx'}, \mathbf{\Sigma}_{x'x}, \mathbf{\Sigma}_{x'x'}$, where the mean blocks are the x- and x'-blocks of $\mathbf{M}_{\ell+1}(\mathbf{m}^\ell, \mathbf{Q}^\ell) = \mathbf{b}_{\ell+1} + \mu_{\ell+1}\mathbf{B}_\ell\mathbf{m}^\ell$ in (15), and the covariance blocks come from $\mathbf{\Sigma}^{\ell+1} = \sigma_{\ell+1}^2\mathbf{B}_\ell\mathbf{Q}^\ell\mathbf{B}_\ell^\top$.

Then

$$\mathbf{K}^{\ell+1}(x,x')_{t,t'} = \sum_{\mathbf{s},\mathbf{s}'\in\{0,1\}^T} s_t s'_{t'}\, \mathbb{P}\left(\mathbf{H}_x \in \mathcal{R}_{\mathbf{s}},\, \mathbf{H}_{x'} \in \mathcal{R}_{\mathbf{s}'}\right). \tag{29}$$

Equation (29) is the temporal hard-SNN analogue of the familiar activation expectation defining a conventional NNGP kernel. The difference is that a stateful spiking neuron maps an entire Gaussian temporal path to an entire spike path, so the kernel recursion acts on $T \times T$ temporal matrices rather than scalar covariances. In particular, leak, reset, refractory structure, and exact spike timing enter the kernel through the geometry of the regions $\{\mathcal{R}_{\mathbf{s}}\}$.

A.1.9 From the temporal kernel to a Gaussian-process prior

The hidden spike process itself is Bernoulli-valued rather than Gaussian. A standard Gaussian linear readout nevertheless converts the limiting temporal kernel into a Gaussian-process prior over functions. Let $\mathbf{c} \in \mathbb{R}^T$ be a fixed temporal readout vector and define $f_n(x) = \sigma_{\mathrm{out}} n_L^{-1/2} \sum_{i=1}^{n_L} a_i\, \mathbf{c}^\top \mathbf{S}_i^L(x)$, where $a_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1)$, with the readout weights independent of all hidden-layer parameters.

Corollary A.6 (SNN Gaussian-process readout)**.** *Under the assumptions of Theorem A.5, $f_n \Rightarrow \mathcal{GP}\left(0, K_{\mathrm{SNN}}\right)$ on every finite input set, where*

$$K_{\mathrm{SNN}}(x,x') = \sigma_{\mathrm{out}}^2 \mathbf{c}^\top \mathbf{K}^L(x,x')\mathbf{c}. \tag{30}$$

Proof. Conditioned on the final-layer spike trajectories, $\left(f_n(x^1), \ldots, f_n(x^M)\right)$ is exactly multivariate Gaussian because the readout coefficients a_i are Gaussian. Its conditional covariance between inputs x^a and x^b is $K_n(x^a,x^b) = \sigma_{\mathrm{out}}^2 n_L^{-1} \sum_{i=1}^{n_L} (\mathbf{c}^\top \mathbf{S}_i^L(x^a))(\mathbf{c}^\top \mathbf{S}_i^L(x^b)) = \sigma_{\mathrm{out}}^2 \mathbf{c}^\top \mathbf{Q}_{n,ab}^L \mathbf{c}$.

By Theorem A.5, $\mathbf{Q}_{n,ab}^L \xrightarrow{p} \mathbf{Q}_{ab}^L$. Hence $K_n(x^a,x^b) \xrightarrow{p} \sigma_{\mathrm{out}}^2 \mathbf{c}^\top \mathbf{Q}_{ab}^L \mathbf{c}$. The conditional characteristic function therefore converges to that of the centered Gaussian distribution with deterministic covariance matrix given by (30). This proves the finite-dimensional Gaussian-process convergence. □

Several conventional SNN readouts are special cases of $\mathbf{c}$. For example, the normalized rate readout corresponds to $\mathbf{c}_{\mathrm{rate}} = \frac{1}{T}\mathbf{1}$, while a last-time readout corresponds to $\mathbf{c}_{\mathrm{last}} = \mathbf{e}_T$. The temporal kernel $\mathbf{K}^L(x,x')$, rather than any particular choice of $\mathbf{c}$, is therefore the architecture-level object that contains the complete second-order temporal geometry available to downstream readouts.

A.1.10 Why second order is sufficient for coarse-graining

At finite width, Proposition A.1 makes $(\mathbf{m}_n^\ell, \mathbf{Q}_n^\ell)$ conditionally sufficient for the next Gaussian mixing layer; Theorem A.5 then makes this descriptor deterministic at infinite width. This second-order coarse-graining does not imply rate sufficiency, because $\mathbf{Q}^\ell$ can retain temporal directions annihilated by rate averaging; the next section asks when the architecture removes those directions.

A.2 Rate-Collapse Proofs and Extended Derivations

Section A.1 established that a wide hard-spiking network is characterized by an operator-valued temporal kernel $\mathbf{K}^\ell(x,x') \in \mathbb{R}^{T\times T}$. We now ask when this temporal object contains information beyond firing rate.

The main idea is to regard rate coding as a quotient of the temporal representation. The constant temporal direction retains the spike count, while all orthogonal directions encode timing information invisible to rate averaging. We show first that if the architecture annihilates these nonconstant directions, the complete SNN function space factors exactly through rate statistics. We then prove an approximate version for hard LIF dynamics. Because the hard threshold is discontinuous, this approximation cannot in general be obtained from a Lipschitz argument. Instead, we couple the full and rate-projected Gaussian fields and control the probability that the perturbation crosses a LIF threshold boundary.

A.2.1 THE RATE QUOTIENT

Recall the normalized constant temporal vector $\mathbf{v}_0 = \frac{1}{\sqrt{T}}\mathbf{1}_T$, $\mathbf{P}_0 = \mathbf{v}_0\mathbf{v}_0^\top$, and define its orthogonal complement $\mathbf{P}_\perp = \mathbf{I}_T - \mathbf{P}_0$. For the stacked representation over M inputs, let $\mathbf{\Pi} = \mathbf{I}_M \otimes \mathbf{P}_0$, $\mathbf{\Pi}_\perp = \mathbf{I}_D - \mathbf{\Pi}$, and $D = MT$.

For a temporal descriptor $c = (\mathbf{m}, \mathbf{Q})$, we define its *rate quotient* $\mathcal{R}(c) := (\mathbf{\Pi m},\ \mathbf{\Pi Q \Pi})$. The first component records the firing rate of each input, replicated along the constant temporal direction. The second records pairwise second moments of those rates.

Indeed, for the (a, b)-th temporal block, $T^{-1}\mathbf{v}_0^\top \mathbf{Q}_{ab}\mathbf{v}_0 = \mathbb{E}[\bar{S}(x^a)\bar{S}(x^b)]$, where $\bar{S}(x) = T^{-1}\sum_{t=1}^{T} S(x, t)$. All components contained in $\mathbf{\Pi}_\perp \mathbf{m}$, $\mathbf{\Pi Q \Pi}_\perp$, $\mathbf{\Pi}_\perp \mathbf{Q \Pi}$, $\mathbf{\Pi}_\perp \mathbf{Q \Pi}_\perp$ are invisible to a rate-only representation.

We therefore call them collectively the *non-DC temporal component*.

A.2.2 EXACT FACTORIZATION THROUGH RATES

Consider one layer with temporal mixing operator $\mathbf{B} = \mathbf{I}_M \otimes \mathbf{A}$. The Gaussian field induced by a descriptor $c = (\mathbf{m}, \mathbf{Q})$ has parameters $\mathbf{M}(c) = \mathbf{b} + \mu\mathbf{Bm}$, and $\mathbf{\Sigma}(c) = \sigma^2\mathbf{BQB}^\top$. The simplest case occurs when the temporal operator removes every nonconstant temporal direction.

Proposition A.7 (Exact rate factorization)**.** *Suppose*

$$\mathbf{B\Pi}_\perp = \mathbf{0}, \tag{31}$$

or equivalently, $\mathbf{AP}_\perp = \mathbf{0}$. *Then the Gaussian preactivation law depends on* $c = (\mathbf{m}, \mathbf{Q})$ *only through its rate quotient* $\mathcal{R}(c)$. *Consequently, there exists a map* $\overline{\mathcal{T}}$ *such that*

$$\mathcal{T}(\mathbf{m}, \mathbf{Q}) = \overline{\mathcal{T}}\left(\mathcal{R}(\mathbf{m}, \mathbf{Q})\right). \tag{32}$$

This statement holds for arbitrary measurable neuron path maps, including hard LIF dynamics.

Proof. Condition (31) implies $\mathbf{B} = \mathbf{B\Pi}$. Hence $\mathbf{Bm} = \mathbf{B\Pi m}$ and $\mathbf{BQB}^\top = \mathbf{B\Pi Q\Pi B}^\top$.

Therefore both the mean and covariance of the Gaussian field are determined completely by $(\mathbf{\Pi m}, \mathbf{\Pi Q \Pi})$. The output spike law is the push-forward of this Gaussian measure through the neuron path map. It therefore also depends only on the rate quotient. Taking its first and second moments yields (32). □

Applied layer by layer, Proposition A.7 immediately yields the following result.

Corollary A.8 (Exact rate-reducible SNN)**.** *Suppose*

$$\mathbf{A}_\ell \mathbf{P}_\perp = \mathbf{0} \qquad \text{for every } \ell = 0, \ldots, L-1. \tag{33}$$

Then the entire infinite-width SNN recursion factors through rate statistics. In particular, if two input spike populations satisfy $\mathcal{R}(c^0) = \mathcal{R}(\widetilde{c}^0)$, *then* $c^\ell = \widetilde{c}^\ell$ *for every* $\ell \geq 1$, *and therefore every Gaussian-readout NNGP kernel agrees:* $K_{\mathrm{SNN}}(x, x') = \widetilde{K}_{\mathrm{SNN}}(x, x')$.

Thus exact rate reducibility is a statement about the information transmitted by the temporal operator, rather than about the presence or absence of spikes.

Condition (33) is intentionally strong. It states that every non-DC temporal direction is removed before reaching the next neuron. The remainder of this section asks what happens when these directions are merely small.

A.2.3 A RATE-PROJECTED SURROGATE

For a general temporal operator $\mathbf{B}$, define its rate-projected counterpart as $\mathbf{B}_{\mathrm{dc}} := \mathbf{B\Pi}$.

Given the same temporal descriptor $c = (\mathbf{m}, \mathbf{Q})$, define the full Gaussian field

$$\mathbf{Z} = \mathbf{b} + \mu\mathbf{Bm} + \sigma\mathbf{BQ}^{1/2}\boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I}_D), \tag{34}$$

and the rate-projected field

$$\mathbf{Z}_{\mathrm{dc}} = \mathbf{b} + \mu\mathbf{B\Pi m} + \sigma\mathbf{B\Pi Q}^{1/2}\boldsymbol{\xi}. \tag{35}$$

The use of the same Gaussian vector $\boldsymbol{\xi}$ gives an explicit coupling of the two fields.

The marginal law of $\mathbf{Z}$ is $\mathcal{N}(\mathbf{b} + \mu\mathbf{B}\mathbf{m}, \sigma^2\mathbf{B}\mathbf{Q}\mathbf{B}^\top)$, whereas $\mathbf{Z}_{\mathrm{dc}} \sim \mathcal{N}(\mathbf{b} + \mu\mathbf{B}\boldsymbol{\Pi}\mathbf{m}, \sigma^2\mathbf{B}\boldsymbol{\Pi}\mathbf{Q}\boldsymbol{\Pi}\mathbf{B}^\top)$. Therefore $\mathbf{Z}_{\mathrm{dc}}$ depends only on the rate quotient $\mathcal{R}(c)$.

We define the associated rate-projected moment map by $\mathcal{T}_{\mathrm{dc}}(c) := \mathcal{M}(\Psi_{\mathrm{LIF}}^{(M)}(\mathbf{Z}_{\mathrm{dc}}))$, where $\mathcal{M}$ returns first and second spike moments.

A.2.4 Temporal leakage energy

The coupling (34)–(35) defines the temporal energy that reaches the neuron. Define $\mathcal{E}_{\mathrm{temp}}(c; \mathbf{B}) := \mu^2\|\mathbf{B}\boldsymbol{\Pi}_\perp\mathbf{m}\|_2^2 + \sigma^2\,\mathrm{tr}(\mathbf{B}\boldsymbol{\Pi}_\perp\mathbf{Q}\boldsymbol{\Pi}_\perp\mathbf{B}^\top)$. This quantity has an exact coupling interpretation.

Lemma A.9 (Temporal leakage as Gaussian perturbation energy)**.** *Under the coupling* (34)–(35)*,*

$$\mathbb{E}\left[\|\mathbf{Z} - \mathbf{Z}_{\mathrm{dc}}\|_2^2\right] = \mathcal{E}_{\mathrm{temp}}(c; \mathbf{B}). \tag{36}$$

Proof. We have $\mathbf{Z} - \mathbf{Z}_{\mathrm{dc}} = \mu\mathbf{B}\boldsymbol{\Pi}_\perp\mathbf{m} + \sigma\mathbf{B}\boldsymbol{\Pi}_\perp\mathbf{Q}^{1/2}\boldsymbol{\xi}$. The second term has zero mean, so the cross term vanishes after expectation. Hence $\mathbb{E}\|\mathbf{Z} - \mathbf{Z}_{\mathrm{dc}}\|_2^2 = \mu^2\|\mathbf{B}\boldsymbol{\Pi}_\perp\mathbf{m}\|_2^2 + \sigma^2\,\mathrm{tr}[\mathbf{B}\boldsymbol{\Pi}_\perp\mathbf{Q}^{1/2}(\mathbf{Q}^{1/2})^\top\boldsymbol{\Pi}_\perp\mathbf{B}^\top] = \mu^2\|\mathbf{B}\boldsymbol{\Pi}_\perp\mathbf{m}\|_2^2 + \sigma^2\,\mathrm{tr}(\mathbf{B}\boldsymbol{\Pi}_\perp\mathbf{Q}\boldsymbol{\Pi}_\perp\mathbf{B}^\top)$. This is (36). □

The quantity $\mathcal{E}_{\mathrm{temp}}$ separates two reasons why an SNN may behave like a rate network: the architecture may suppress temporal directions, $\|\mathbf{B}\boldsymbol{\Pi}_\perp\|_{\mathrm{op}} \ll 1$, or the input population may contain little energy in them. Rate reducibility is therefore controlled by temporal content together with temporal transmission, not either factor alone. Because every realizable spike descriptor satisfies $\|\mathbf{m}\|_2^2 \le D$ and $\mathrm{tr}(\mathbf{Q}) \le D$, the condition $\|\mathbf{B}\boldsymbol{\Pi}_\perp\|_{\mathrm{op}} \le \varepsilon$ implies

$$\mathcal{E}_{\mathrm{temp}}(c; \mathbf{B}) \le D(\mu^2 + \sigma^2)\varepsilon^2. \tag{37}$$

A.2.5 Why hard thresholds require anti-concentration

For a smooth activation, a small perturbation $\mathbf{Z} - \mathbf{Z}_{\mathrm{dc}}$ could be propagated directly by a Lipschitz inequality. The hard-LIF map does not satisfy such an inequality. Nevertheless, the spike pattern can change only if the perturbation crosses a boundary of the polyhedral LIF partition.

Let $\mathfrak{B}_{\mathrm{LIF}} := \bigcup_{\mathbf{s}} \partial\mathcal{R}_{\mathbf{s}}$ denote the union of all threshold boundaries over the M input blocks. By Lemma A.2, $\mathfrak{B}_{\mathrm{LIF}}$ is contained in a finite union of affine hyperplanes. We use the following local anti-concentration assumption.

Assumption 2 (LIF boundary anti-concentration). For the rate-projected Gaussian field $\mathbf{Z}_{\mathrm{dc}}$, there exist constants $C_{\mathrm{ac}} > 0$ and $r_0 > 0$ such that $\mathbb{P}[\mathrm{dist}(\mathbf{Z}_{\mathrm{dc}}, \mathfrak{B}_{\mathrm{LIF}}) \le r] \le C_{\mathrm{ac}} r$ for all $0 < r \le r_0$.

Assumption 2 is weaker than requiring the Gaussian covariance to be nonsingular. The rate-projected Gaussian field may live on a lower-dimensional subspace. It is sufficient that this subspace intersects every relevant threshold hyperplane transversely, or stays a positive distance away from it.

For example, suppose $\mathfrak{B}_{\mathrm{LIF}} \subseteq \bigcup_{q=1}^{J}\{\mathbf{z} : \mathbf{a}_q^\top\mathbf{z} = b_q\}$, $\|\mathbf{a}_q\|_2 = 1$. By Lemma A.2 one may take $J = MT2^{T-1}$: writing $\mathbf{z}_a \in \mathbb{R}^T$ for the a-th input block of $\mathbf{z}$, the boundary lies in the hyperplanes $\{\mathbf{z} : (\mathbf{L}_\lambda\mathbf{z}_a)_t = c_t(\mathbf{s})\}$, rescaled to unit normal, and the offset $c_t(\mathbf{s})$ of (14) depends only on $s_1, \ldots, s_{t-1}$, so it takes at most $2^{t-1} \le 2^{T-1}$ values for each (a, t).

If every relevant boundary satisfies $\mathrm{Var}(\mathbf{a}_q^\top\mathbf{Z}_{\mathrm{dc}}) \ge \kappa^2 > 0$, then the one-dimensional Gaussian density bound gives

$$C_{\mathrm{ac}} \le \frac{2J}{\sqrt{2\pi}\kappa}. \tag{38}$$

Lemma A.10 (Boundary-crossing stability)**.** *Let* $\mathbf{Z}$ *and* $\mathbf{Z}_0$ *be random vectors defined on the same probability space. Suppose* $\mathbf{Z}_0$ *satisfies Assumption 2 and* $\mathbb{E}\|\mathbf{Z} - \mathbf{Z}_0\|_2^2 \le \delta^2$*. Then, whenever the optimizing radius below is at most* r_0*,*

$$\mathbb{P}\left[\Psi_{\mathrm{LIF}}^{(M)}(\mathbf{Z}) \ne \Psi_{\mathrm{LIF}}^{(M)}(\mathbf{Z}_0)\right] \le C_\star C_{\mathrm{ac}}^{2/3}\delta^{2/3}, \tag{39}$$

where $C_\star = 3 \cdot 2^{-2/3}$.

Proof. If $\Psi_{\mathrm{LIF}}^{(M)}(\mathbf{Z}) \neq \Psi_{\mathrm{LIF}}^{(M)}(\mathbf{Z}_0)$, then $\mathbf{Z}$ and $\mathbf{Z}_0$ belong to different cells of the polyhedral partition. The line segment connecting them must therefore intersect $\mathfrak{B}_{\mathrm{LIF}}$. Consequently, $\mathrm{dist}(\mathbf{Z}_0, \mathfrak{B}_{\mathrm{LIF}}) \leq \|\mathbf{Z} - \mathbf{Z}_0\|_2$. For any $0 < r \leq r_0$, $\mathbb{P}[\Psi(\mathbf{Z}) \neq \Psi(\mathbf{Z}_0)] \leq \mathbb{P}[\mathrm{dist}(\mathbf{Z}_0, \mathfrak{B}_{\mathrm{LIF}}) \leq r] + \mathbb{P}[\|\mathbf{Z} - \mathbf{Z}_0\|_2 > r] \leq C_{\mathrm{ac}} r + \delta^2/r^2$, where the second term follows from Markov's inequality applied to $\|\mathbf{Z} - \mathbf{Z}_0\|_2^2$.

Choosing $r = \left(\frac{2\delta^2}{C_{\mathrm{ac}}}\right)^{1/3}$, which is at most r_0 by hypothesis (when $\delta = 0$, $\mathbf{Z} = \mathbf{Z}_0$ almost surely and (39) holds trivially, so take $\delta > 0$), gives $C_{\mathrm{ac}} r + \frac{\delta^2}{r^2} = 3 \cdot 2^{-2/3} C_{\mathrm{ac}}^{2/3} \delta^{2/3}$. This proves (39). □

The exponent $2/3$ is not intended to be optimal. It results from combining a linear Gaussian anti-concentration estimate with only a second-moment perturbation bound. Sharper sub-Gaussian tail control yields correspondingly sharper moduli. Thus, the hard threshold is stable in probability even though it is not pointwise Lipschitz.

A.2.6 One-layer approximate rate collapse

We can now quantify the loss incurred by discarding non-DC temporal information.

Theorem A.11 (One-layer hard-LIF rate collapse)**.** *Let $c = (\mathbf{m}, \mathbf{Q})$ be a realizable temporal descriptor and let $\mathcal{T}(c)$ and $\mathcal{T}_{\mathrm{dc}}(c)$ denote the first and second output spike moments generated respectively by $\mathbf{Z}$ and $\mathbf{Z}_{\mathrm{dc}}$ in* (34)–(35)*.*

Suppose Assumption 2 holds for $\mathbf{Z}_{\mathrm{dc}}$, and the optimizing radius $(2\mathcal{E}_{\mathrm{temp}}/C_{\mathrm{ac}})^{1/3}$ is at most r_0. Otherwise use the infimum bound (6)*. Then*

$$\|\mathcal{T}(c) - \mathcal{T}_{\mathrm{dc}}(c)\|_\infty \leq C_\star C_{\mathrm{ac}}^{2/3} \mathcal{E}_{\mathrm{temp}}(c; \mathbf{B})^{1/3}. \tag{40}$$

In particular, if $\|\mathbf{B}\mathbf{\Pi}_\perp\|_{\mathrm{op}} \leq \varepsilon$, then

$$\|\mathcal{T}(c) - \mathcal{T}_{\mathrm{dc}}(c)\|_\infty \leq C\varepsilon^{2/3}, \tag{41}$$

where $C = C_\star C_{\mathrm{ac}}^{2/3} \left[D(\mu^2 + \sigma^2)\right]^{1/3}$.

Proof. Apply Lemma A.10 to the coupling $(\mathbf{Z}, \mathbf{Z}_{\mathrm{dc}})$. By Lemma A.9, $\delta^2 = \mathcal{E}_{\mathrm{temp}}(c; \mathbf{B})$. Let $\mathbf{S} = \Psi_{\mathrm{LIF}}^{(M)}(\mathbf{Z})$, $\mathbf{S}_{\mathrm{dc}} = \Psi_{\mathrm{LIF}}^{(M)}(\mathbf{Z}_{\mathrm{dc}})$. Every coordinate of $\mathbf{S}$ belongs to $\{0, 1\}$, hence $|S_r - S_{\mathrm{dc},r}| \leq \mathbf{1}\{\mathbf{S} \neq \mathbf{S}_{\mathrm{dc}}\}$. Similarly, $|S_r S_q - S_{\mathrm{dc},r} S_{\mathrm{dc},q}| \leq \mathbf{1}\{\mathbf{S} \neq \mathbf{S}_{\mathrm{dc}}\}$. Therefore every coordinate of the difference between the first and second moment descriptors is bounded by the mismatch probability. Equation (40) follows from (39).

Equation (41) follows from (37). □

Theorem A.11 is the hard-threshold analogue of a conventional activation-perturbation bound. Its controlling quantity, $\mathcal{E}_{\mathrm{temp}}(c; \mathbf{B})$, has a direct interpretation: it is exactly the mean-squared portion of the Gaussian synaptic field contributed by temporal directions that rate coding discards.

A.2.7 Observed local scaling

The general result above gives an $\mathcal{E}_{\mathrm{temp}}^{1/3}$ bound. The tested local perturbation has a smaller, nearly linear gap, but we do not promote that observation to a general crossing theorem: each rate-projected input block has only one Gaussian degree of freedom, so after conditioning on a nondegenerate prefix its later innovation variances vanish. Taking $\mathbf{B} = \mathbf{\Pi} + \varepsilon\mathbf{\Pi}_\perp$ at $T = 4$, so that $\|\mathbf{B}\mathbf{\Pi}_\perp\|_{\mathrm{op}} = \varepsilon$, gives Figure 10.

The first curve is nearly constant over a factor of 64 in the tested perturbation scale, while the second decreases. This is consistent with a linear gap in this experiment; it does not prove a universal exponent. The experiment takes $M = 1$, $\lambda = \rho = \vartheta = \mu = 0$, $\mathbf{b} = \mathbf{0}$, $\sigma = 1$ and a realizable symmetric Toeplitz second moment with first row $(0.5, 0.3, 0.15, 0.05)$. This positive rescaling leaves the zero-threshold sign-output curve unchanged, and (7) gives its output moments in closed form.

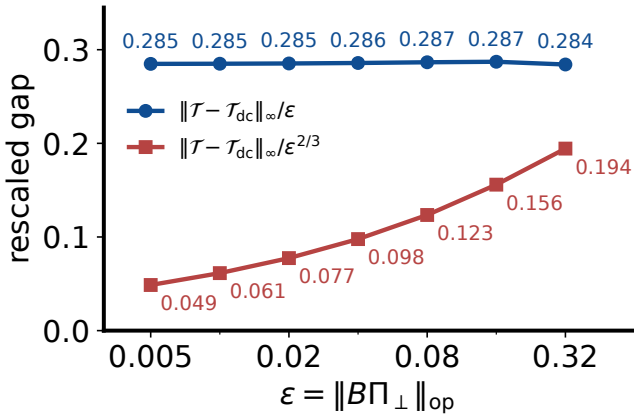


Figure 10: One-layer gap $\|\mathcal{T} - \mathcal{T}_{\mathrm{dc}}\|_\infty$ divided by ε and by $\varepsilon^{2/3}$, at $T = 4$.

A.2.8 Finite-depth kernel collapse

We next propagate the one-layer result through a finite-depth network.

The projected map is discontinuous at some realizable descriptors: for $M = T = 1$, $\mu = \vartheta = 0$, $\mathbf{b} = \mathbf{0}$, zero initial membrane value, $\sigma = 1$, and $\mathbf{B} = \mathbf{I}_1$, its first output moment is $1/2$ when the scalar entry $Q_{11} > 0$ but 1 when $Q_{11} = 0$. A modulus taken over all realizable descriptors therefore need not vanish. The following theorem uses compact admissible sets and a common vanishing modulus to exclude this failure.

Let $c^{\ell+1} = \mathcal{T}_\ell(c^\ell)$ denote the full temporal recursion.

Define the rate-projected recursion by

$$\bar{c}^{\ell+1} = \mathcal{T}_{\ell,\mathrm{dc}}(\bar{c}^\ell), \qquad \bar{c}^0 = c^0. \tag{42}$$

Although $\bar{c}^\ell$ itself contains a full spike-path moment matrix generated by the LIF dynamics, the information transmitted through every subsequent layer depends only on $\mathcal{R}(\bar{c}^\ell)$. Thus (42) defines a genuine rate-quotient computation.

Let $\mathfrak{C}_D = \{(\mathbf{m}, \mathbf{Q}) : \exists\, \nu \in \mathcal{P}(\{0,1\}^D),\ \mathbf{m} = \mathbb{E}_\nu \mathbf{S},\ \mathbf{Q} = \mathbb{E}_\nu \mathbf{S}\mathbf{S}^\top\}$ denote the set of realizable spike descriptors. Because the spike-path space is finite, $\mathfrak{C}_D$ is compact.

For each layer choose a compact admissible subset $\mathfrak{C}^{\mathrm{adm}}_{D,\ell} \subseteq \mathfrak{C}_D$ containing both descriptors being compared, on which every projected Gaussian field is boundary-null. For parameter families assume a common modulus of continuity on these sets, uniform in the family. Pointwise boundary-nullness only at one limiting descriptor is insufficient. For each projected layer map define its modulus of continuity by $\omega_\ell(r) := \sup\{\|\mathcal{T}_{\ell,\mathrm{dc}}(c) - \mathcal{T}_{\ell,\mathrm{dc}}(\widetilde{c})\|_\infty : c, \widetilde{c} \in \mathfrak{C}^{\mathrm{adm}}_{D,\ell},\ \|c - \widetilde{c}\|_\infty \le r\}$.

Under the boundary-null condition on the compact admissible set just specified, $\omega_\ell(r) \to 0$ as $r \downarrow 0$.

Theorem A.12 (Finite-depth rate-kernel collapse)**.** *Suppose Assumption 2 holds uniformly along the full and rate-projected limiting trajectories, the optimizing radii are admissible, and the projected maps have the common vanishing moduli on the compact admissible sets specified above. Define* $e_\ell = \|c^\ell - \bar{c}^\ell\|_\infty$ *and* $\eta_\ell = C_\star C^{2/3}_{\mathrm{ac},\ell} \mathcal{E}_{\mathrm{temp}}(c^\ell; \mathbf{B}_\ell)^{1/3}$.

Then

$$e_{\ell+1} \le \eta_\ell + \omega_\ell(e_\ell). \tag{43}$$

Consequently, for any fixed depth L, if $\max_{\ell<L} \mathcal{E}_{\mathrm{temp}}(c^\ell; \mathbf{B}_\ell) \longrightarrow 0$, *then*

$$\|c^L - \bar{c}^L\|_\infty \longrightarrow 0. \tag{44}$$

For any fixed temporal readout vector $\mathbf{a}$, $K_{\mathrm{SNN}}(x, x') = \sigma^2_{\mathrm{out}} \mathbf{a}^\top \mathbf{Q}^L_{xx'} \mathbf{a}$ *and its rate-quotient counterpart* $\bar{K}_{\mathrm{SNN}}$ *therefore satisfy* $\sup_{x,x' \in \mathcal{X}_M} |K_{\mathrm{SNN}}(x, x') - \bar{K}_{\mathrm{SNN}}(x, x')| \longrightarrow 0$.

Proof. Insert an intermediate projected update evaluated at the full descriptor: $e_{\ell+1} = \|\mathcal{T}_\ell(c^\ell) - \mathcal{T}_{\ell,\mathrm{dc}}(\bar{c}^\ell)\|_\infty \le \|\mathcal{T}_\ell(c^\ell) - \mathcal{T}_{\ell,\mathrm{dc}}(c^\ell)\|_\infty + \|\mathcal{T}_{\ell,\mathrm{dc}}(c^\ell) - \mathcal{T}_{\ell,\mathrm{dc}}(\bar{c}^\ell)\|_\infty$.

The first term is at most η_ℓ by Theorem A.11. The second is at most $\omega_\ell(e_\ell)$ by definition. This proves (43).

Since $e_0 = 0$, finite induction together with $\eta_\ell \to 0$ and $\omega_\ell(r) \to 0$ as $r \to 0$ gives (44). Finally, $|K_{\mathrm{SNN}}(x,x') - \bar{K}_{\mathrm{SNN}}(x,x')| \le \sigma_{\mathrm{out}}^2 \|a\|_1^2 \|\mathbf{Q}^L_{xx'} - \bar{\mathbf{Q}}^L_{xx'}\|_\infty$, which tends to zero by (44). □

Theorem A.12 therefore turns vanishing transmitted non-DC energy into a vanishing NNGP kernel gap, without requiring the hard-LIF map itself to become smooth.

A.2.9 Architecture-level sufficient condition

Corollary A.13 (Architecture-level approximate collapse)**.** *Suppose* $\max_{\ell<L} \|\mathbf{B}_\ell \mathbf{\Pi}_\perp\|_{\mathrm{op}} \le \varepsilon$*, while depth, horizon, parameters, anti-concentration constants and the common moduli in Theorem A.12 remain uniformly bounded. Then* $\mathcal{E}_{\mathrm{temp}}(c^\ell; \mathbf{B}_\ell) = O(\varepsilon^2)$ *uniformly over realizable descriptors, and at fixed depth the network and rate-projected kernels converge as* $\varepsilon \to 0$*. At one layer,* $\|\mathcal{T}_\ell - \mathcal{T}_{\ell,\mathrm{dc}}\|_\infty = O(\varepsilon^{2/3})$.

The condition is distribution free and sufficient, not necessary: a particular input law can have negligible temporal leakage even when the operator can transmit temporal modes.

A.2.10 Depth diagnostics and conditional contraction

For a centered memoryless zero-threshold layer with $\mathbf{B} = \mathbf{I}$, unit-diagonal normalization and Price's Gaussian identity (Price, 1958) give $g(r) = \frac{1}{2} + \arcsin(r)/\pi, qquad g'(r) = 1/(\pi\sqrt{1-r^2})$. Its interior fixed point is 0.789833, with derivative 0.5190; the derivative diverges near $r = 1$, so this does not prove global contraction.

Finite sweep. At $T = 4$, we evaluate all 280 points in an $8 \times 5 \times 7$ grid of leak, reset and threshold values using 2^{20} shared quasi-random samples, five starts and unit-diagonal renormalization. Jacobian spectral radii have minimum 0.082, median 0.348 and maximum 0.550; exact Gaussian-orthant evaluation of the leading band gives maximum 0.5513. Eighty additional combinations of ten temporal operators and eight neuron settings give maximum 0.555. Every tested configuration contracts locally, but the finite sweep is not a universal bound over all operators or parameters.

Conditional encoding depth. On any invariant set with contraction constant $0 < \chi < 1$, two trajectories of the same map satisfy $\|\mathbf{Q}_P^L - \mathbf{Q}_Q^L\| \le C\chi^L$, giving $L^*(\tau) = \log(\tau/C)/\log\chi$. For the exact memoryless map, the witness gap falls from 0.108 after one layer to 4.1×10^{-5} after twelve, with ratio approaching 0.519. This statement concerns two inputs under one map. A full network and its projected surrogate can approach different fixed points; depth alone therefore does not imply rate collapse.

A.2.11 Stationary kernels and temporal mass

For a stationary temporal kernel on a finite periodic window, the Fourier basis diagonalizes $\mathbf{K}$, and $\mathbf{P}_0$ retains its zero-frequency component. This motivates

$$\mathfrak{M}_{\mathrm{temp}}(\mathbf{K}) = \frac{\mathrm{tr}(\mathbf{P}_\perp \mathbf{K} \mathbf{P}_\perp)}{\mathrm{tr}(\mathbf{K})} = \frac{\sum_{\omega \ne 0} \lambda_\omega}{\sum_\omega \lambda_\omega}. \tag{45}$$

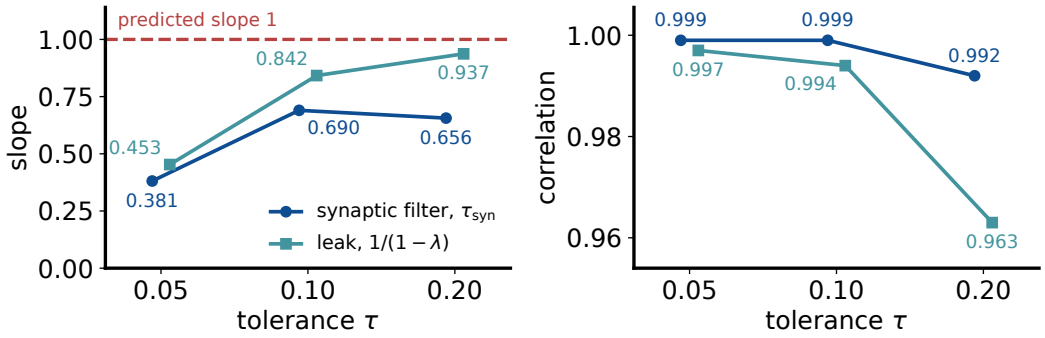
Nonzero mass certifies temporal structure, not its downstream use; the operator and spike map may still suppress it.

A.2.12 Coarse-graining at finite temporal resolution

Let $\mathbf{P}_\Delta$ project onto signals constant on contiguous blocks of length Δ, and $\mathbf{\Pi}_\Delta = \mathbf{I}_M \otimes \mathbf{P}_\Delta$. Then $\Delta = 1$ is the identity and $\Delta = T$ is the rate projector. The collapse results extend after replacing $\mathbf{\Pi}$ by $\mathbf{\Pi}_\Delta$, with exact factorization when $\mathbf{A}_\ell \mathbf{P}_{\Delta,\perp} = 0$.

Definition A.14 (Encoding timescale)**.** For tolerance $\tau > 0$, define $\Delta^*(\tau) = \max\{\Delta : \|K_{\mathrm{SNN}} - K_{\mathbf{\Pi}_\Delta}\| \le \tau\}$, the coarsest block resolution that leaves the induced kernel within tolerance.

Timescale check. At depth two and $T = 16$, we sweep exponential-filter constants $\tau_{\text{syn}} \in \{1, 2, 4, 8\}$, leak values $\lambda \in \{0, 0.3, 0.5, 0.7, 0.8, 0.9\}$, block sizes $\Delta \in \{1, 2, 4, \ldots, T\}$ and tolerances 0.05–0.3, recalibrating all arms to firing rate 0.25. The gap is zero at $\Delta = 1$ and monotone in every run. Correlations of Δ^* with τ_{syn} and $1/(1-\lambda)$ range from 0.895 to 1.000, but fitted slopes are below one and tolerance dependent. At $T = 32$, the filter-family exponent is 0.71–0.74 across seven realizations, with correlation at least 0.998. Thus Δ^* tracks and orders the tested architectural timescales but is not equal to them.



(a) log–log slope (b) correlation

Figure 11: Slope and correlation of Δ^* against filter and leak timescales.

On the same finite periodic window, the block projector is not circulant. Its weight on Fourier mode f_ω is the Fejér factor $\|\mathbf{P}_\Delta f_\omega\|^2 = \sin^2(\omega\Delta/2)/(\Delta^2 \sin^2(\omega/2))$, so block averaging is a soft low pass rather than a sharp frequency-band cutoff. Finally, embedding the separating pair in one block shows that temporal irreducibility can persist at every $2 \le \Delta \le T$; the rate-versus-timing distinction is therefore a hierarchy over temporal resolution.

A.2.13 Discussion: what the collapse theorem does and does not say

Raw temporal variation is not the relevant quantity: it matters only to the extent that the synaptic operator transmits it. Conversely, nonzero transmitted leakage need not be observable if the downstream threshold or readout is insensitive to that direction. The next section therefore gives a constructive witness in which non-DC structure changes the induced function-space kernel.

A.3 Irreducibility Proofs, Examples, and Readout Extensions

Section A.2 showed when an architecture removes non-DC information. We now ask for the converse: can rate-equivalent inputs induce different downstream kernels? For hard thresholds this question is decidable through Gaussian orthant probabilities.

A.3.1 Rate equivalence and temporal irreducibility

Let $\mathcal{S}_T = \{0, 1\}^T$ denote the space of finite-horizon spike trajectories, and define the normalized spike count $R(\mathbf{s}) = \frac{1}{T}\mathbf{1}^\top \mathbf{s}$. For a probability distribution $\nu \in \mathcal{P}(\mathcal{S}_T)$, let $R_\# \nu$ denote the push-forward distribution of the firing rate under R.

Definition A.15 (Rate equivalence)**.** Two spike-path distributions $\nu, \widetilde{\nu} \in \mathcal{P}(\mathcal{S}_T)$ are *rate equivalent*, written $\nu \equiv_{\text{rate}} \widetilde{\nu}$, if $R_\# \nu = R_\# \widetilde{\nu}$. Equivalently, the complete probability distribution of their spike count is identical.

Rate equivalence is stronger than equality of mean firing rates. If $\nu \equiv_{\text{rate}} \widetilde{\nu}$, then for every measurable function $g : [0, 1] \to \mathbb{R}$, $\mathbb{E}_\nu[g(R(\mathbf{S}))] = \mathbb{E}_{\widetilde{\nu}}[g(R(\mathbf{S}))]$. Thus no representation, statistic, or predictor that observes a spike trajectory only through its firing rate can distinguish the two populations at the input.

For a wide spiking layer with temporal kernel map $\mathcal{K}$, we say that rate information is sufficient if this equivalence is respected by the downstream function space.

Definition A.16 (Rate reducibility)**.** A temporal kernel map $\mathcal{K}$ is *rate reducible* on a class $\mathcal{P} \subseteq \mathcal{P}(\mathcal{S}_T)$ if $\nu \equiv_{\text{rate}} \widetilde{\nu}$ implies $\mathcal{K}(\nu) = \mathcal{K}(\widetilde{\nu})$ for every $\nu, \widetilde{\nu} \in \mathcal{P}$. Otherwise, $\mathcal{K}$ is *temporally irreducible*.

This definition deliberately concerns the downstream function-space object, rather than the internal presence of temporal variation. A model may contain substantial spike timing variability while remaining rate reducible if its dynamics never convert that variability into a downstream kernel difference. Our goal is therefore to exhibit a pair $\nu \equiv_{\mathrm{rate}} \widetilde{\nu}$ for which $\mathcal{K}(\nu) \neq \mathcal{K}(\widetilde{\nu})$.

A.3.2 Three notions of rate reducibility

Definition A.16 is one of three inequivalent statements that the phrase "rate reducible" can denote, and separating them resolves an apparent conflict between our results and rate matching in ANN-to-SNN conversion. Write $\mathcal{R}(\mathbf{m}, \mathbf{Q}) = (\mathbf{\Pi m}, \mathbf{\Pi Q \Pi})$ for the rate quotient.

- **(R1)** *Descriptor factorisation*: $\mathcal{T}_\ell = \overline{\mathcal{T}}_\ell \circ \mathcal{R}$. This is the content of Proposition A.7 and Theorem 3.3.
- **(R2)** *Quotient closure*: $\mathcal{R} \circ \mathcal{T}_\ell = \widehat{\mathcal{T}}_\ell \circ \mathcal{R}$.
- **(R3)** *Count-law reducibility*: Definition A.16.

Clearly $\mathrm{R1} \Rightarrow \mathrm{R2}$, and R2 holding at every layer already forces the network's rate-readout kernel to be a function of $\mathcal{R}(c^0)$ alone, by induction on depth. The converse fails, and the failure is the point: R1 constrains the *synaptic operator*, whereas R2 can hold for reasons internal to the *neuron*.

The canonical instance is the integrate-and-fire neuron used in ANN-to-SNN conversion (Rueckauer et al., 2017; Bu et al., 2022).

Proposition A.17 (Integrator endpoint)**.** *Let $\lambda = 1$ with soft reset $\rho = \vartheta > 0$, and suppose the temporal operator is DC invariant, $\mathbf{A}^\top \mathbf{1}_T = \gamma \mathbf{1}_T$ (automatic for circulant $\mathbf{A}$). Writing $\widehat{H}_t = \sum_{q \le t} h_q$ and $N_t = \sum_{q=1}^{t} s_{q-1}$ for the reset accumulator — the number of spikes that have fed back by time t, which with $s_0 = 0$ is the count of the first $t-1$ emitted spikes — the exact identity*

$$\vartheta N_T = \widehat{H}_T + u_0 - u_T \tag{46}$$

gives

$$\left| \frac{N_T}{T} - \mathrm{clip}_{[0,1]} \left(\frac{\mathbf{1}^\top \mathbf{H}}{\vartheta T} \right) \right| \le \frac{|u_T - u_0|}{\vartheta T}, \tag{47}$$

because $N_T/T \in [0,1]$ and clipping to $[0,1]$ is 1-Lipschitz; when $\mathbf{1}^\top \mathbf{H}/(\vartheta T) \in [0,1]$ the clip is inactive and the difference inside the absolute value equals $-(u_T - u_0)/(\vartheta T)$ exactly. So the deviation from a function of the DC component alone is at most the terminal residual potential divided by ϑT.

Identity (46) alone does not imply asymptotic closure. For an input family x, suppose instead that its terminal residual is uniformly controlled, $\sup_T \mathbb{E}|u_T(x) - u_0| \le C_x$. With the actual emitted rate $R_T = (N_T + s_T)/T$ and $r_{\mathrm{clip}} = \mathrm{clip}_{[0,1]}(\mathbf{1}^\top \mathbf{H}/(\vartheta T))$, $\mathbb{E}|R_T - r_{\mathrm{clip}}| \le (C_x/\vartheta + 1)/T$. Because both rates lie in $[0,1]$, the corresponding cross-input second moment differs from the clipped-rate moment by at most the sum of the two first-order errors. Thus R2 closes at the rate readout at $O(1/T)$ under uniform terminal-residual control; a multilayer conclusion requires this control at every layer.

The residual condition is not implied by the average drive lying in $(0, \vartheta)$. For $T = 4k$, $\mathbf{A} = \mathbf{I}$, $\lambda = \rho = \vartheta = 1$, $\mu = 2$, $\sigma = 1$, and zero bias and initial state, a spike every fourth bin and k spikes placed in the final k bins have identical counts. Their Gaussian drive is Gs_t with $G \sim \mathcal{N}(2,1)$, but their limiting rates are respectively $\mathrm{clip}_{[0,1]}(G/4)$ and $\mathrm{clip}_{[0,1]}(G)/4$; the rate self-kernel gap tends to 0.25110, rather than zero.

For the tested input families, the following finite-horizon experiment measures when the residual-controlled scaling is observed. Taking two drives $\mathbf{H} \sim \mathcal{N}(\mathbf{m}, \mathbf{I}_T)$ with the same time-averaged mean profile but different non-DC part, $\mathbf{m} = \mathbf{1}_T$ against $\mathbf{m}$ equal to 1.6 on the first $T/2$ bins and 0.4 on the rest, starting from zero membrane potential, and recalibrating the threshold on the first drive to a fixed firing rate of 0.2 at each T, by sixty bisection steps on the same 2^{17} quasi-random Gaussian samples of a single scrambling on which both drives are then evaluated, the rate-readout gap, here the absolute difference of the two mean output firing rates over the T bins and these samples, scaled by T behaves as in Figure 12.

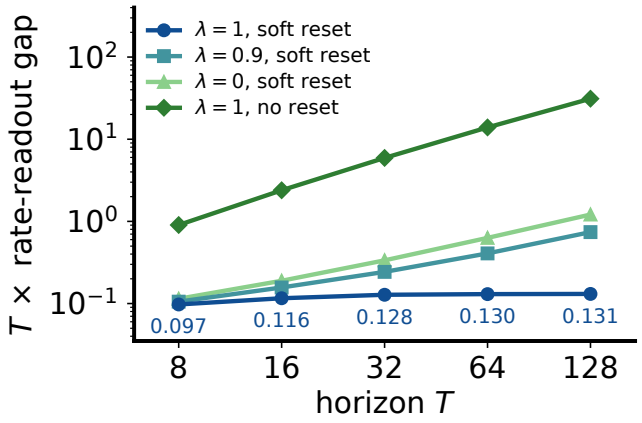


Figure 12: Rate-readout gap times T against T, by leak and reset.

Only the first curve is constant, that is, only there is the gap $\Theta(1/T)$. The descriptor gap $\max|\Delta\mathbf{Q}_{\text{out}}|$ meanwhile stays flat along that curve, at 0.216 to 0.249. For input families satisfying the uniform residual condition, the limiting quotient map is the NNGP of a clipped-ReLU network. This gives a conditional conversion endpoint that is rate reducible at the rate readout but not at the full temporal descriptor.

Two remarks follow. Identity (46) is exact for arbitrary drives; sign changes and the one-spike-per-step ceiling matter only to whether the terminal residual vanishes after normalization. This is the "unevenness error" identified by Bu et al. (2022), expressed in the present notation. It also explains why a timing code can implement a rate computation exactly: the time-to-first-spike construction of Stanojevic et al. (2023) realises a ReLU network without loss of performance, so the use of spike timing does not imply temporal irreducibility. And more generally, rate reducibility is not a property of a network but of a pair, network and readout: the same dynamics that are reducible for $\mathbf{a} \propto \mathbf{1}$ are irreducible for the full temporal kernel. This is the function-space counterpart of the distinction between information that is present in a representation and information that a downstream stage actually uses.

A.3.3 Hard thresholds detect temporal Gaussian correlation

We begin with a general property of a zero-threshold hard-spiking layer.

Consider a centered Gaussian temporal field $\mathbf{H} = (H_1, \dots, H_T)^\top \sim \mathcal{N}(0, \mathbf{\Sigma})$, and the memoryless hard-spike map

$$Y_t = \mathbf{1}\{H_t > 0\}. \tag{48}$$

This is the $\lambda = \rho = \vartheta = 0$ special case of the hard-LIF family.

For any pair $t \neq t'$, define $\rho_{tt'} = \frac{\Sigma_{tt'}}{\sqrt{\Sigma_{tt}\Sigma_{t't'}}}$, assuming $\Sigma_{tt}, \Sigma_{t't'} > 0$.

The classical bivariate Gaussian orthant identity gives $\mathbb{P}(H_t > 0, H_{t'} > 0) = \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho_{tt'})$. Consequently,

$$\mathbb{E}[Y_t Y_{t'}] = \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho_{tt'}). \tag{49}$$

The hard threshold therefore converts Gaussian correlation directly into a spike-kernel entry.

Proposition A.18 (Gaussian orthant witness)**.** *Let two temporal descriptors produce centered Gaussian fields* $\mathbf{H} \sim \mathcal{N}(0, \mathbf{\Sigma})$, $\widetilde{\mathbf{H}} \sim \mathcal{N}(0, \widetilde{\mathbf{\Sigma}})$. *Suppose for some pair* $t \neq t'$, $\rho_{tt'} \neq \widetilde{\rho}_{tt'}$. *Then after the hard-spike map* (48)*,*

$$\left|Q_{\text{out}}(t,t') - \widetilde{Q}_{\text{out}}(t,t')\right| = \frac{1}{2\pi}\left|\arcsin(\rho_{tt'}) - \arcsin(\widetilde{\rho}_{tt'})\right| > 0. \tag{50}$$

In particular,

$$\|\mathbf{Q}_{\text{out}} - \widetilde{\mathbf{Q}}_{\text{out}}\|_{\text{op}} \geq \frac{1}{2\pi}\left|\arcsin(\rho_{tt'}) - \arcsin(\widetilde{\rho}_{tt'})\right|. \tag{51}$$

Proof. Equation (50) follows immediately from (49). For any matrix $\mathbf{A}$, $\|\mathbf{A}\|_{\mathrm{op}} \geq |A_{tt'}|$, which gives (51). □

Proposition A.18 identifies the mechanism directly: Gaussian mixing turns temporal covariance into $\rho_{tt'}$, and the hard threshold turns that correlation into $Q_{\mathrm{out}}(t, t')$. The remaining question is whether two spike populations can differ in this quantity while being completely indistinguishable by rate coding. The answer is yes.

A.3.4 A SAME-RATE, DIFFERENT-KERNEL CONSTRUCTION

A minimal construction already exists for $T = 2$. Let P and Q be distributions on $\{(0,0),(1,0),(0,1),(1,1)\}$ given by $P = (0.1, 0.7, 0.1, 0.1)$ and $Q = (0.1, 0.4, 0.4, 0.1)$, in the order listed. These are finite-horizon spike-path distributions; no stationarity assumption is required.

Their spike-count distributions are identical: $\mathbb{P}_P(\#\mathbf{S} = 0) = \mathbb{P}_Q(\#\mathbf{S} = 0) = 0.1$, $\mathbb{P}_P(\#\mathbf{S} = 1) = \mathbb{P}_Q(\#\mathbf{S} = 1) = 0.8$, and $\mathbb{P}_P(\#\mathbf{S} = 2) = \mathbb{P}_Q(\#\mathbf{S} = 2) = 0.1$. Therefore $P \equiv_{\mathrm{rate}} Q$. In particular, every moment and every nonlinear statistic of the firing rate agrees under P and Q.

Their first temporal moments are nevertheless $\mathbf{m}_P = (0.8, 0.2)^\top$ and $\mathbf{m}_Q = (0.5, 0.5)^\top$, and their second temporal moments are

$$\mathbf{Q}_P = \begin{pmatrix} 0.8 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}, \qquad \mathbf{Q}_Q = \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}. \tag{52}$$

Let $\mathbf{v}_0 = 2^{-1/2}(1,1)^\top$ and $\mathbf{P}_0 = \mathbf{v}_0\mathbf{v}_0^\top$.

One verifies directly that $\mathbf{P}_0\mathbf{m}_P = \mathbf{P}_0\mathbf{m}_Q$, and $\mathbf{P}_0\mathbf{Q}_P\mathbf{P}_0 = \mathbf{P}_0\mathbf{Q}_Q\mathbf{P}_0$. Thus the two distributions agree not only in their complete rate law but also in the first- and second-order DC descriptor used in Section A.2.1.

Now consider a single wide Gaussian layer with

$$\mu = 0, \qquad \mathbf{b} = \mathbf{0}, \qquad \mathbf{A} = \mathbf{I}_2, \qquad \sigma > 0. \tag{53}$$

By Proposition A.1, whose conditional covariance $\sigma^2\mathbf{Q}_n$ converges to $\sigma^2\mathbf{Q}_P$ or $\sigma^2\mathbf{Q}_Q$ by the law of large numbers, the limiting preactivation fields satisfy

$$\mathbf{H}_P \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q}_P), \tag{54}$$

and

$$\mathbf{H}_Q \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q}_Q). \tag{55}$$

The corresponding normalized temporal correlations are $\rho_P = \frac{0.1}{\sqrt{0.8 \cdot 0.2}} = \frac{1}{4}$, and $\rho_Q = \frac{0.1}{\sqrt{0.5 \cdot 0.5}} = \frac{1}{5}$. After the hard threshold $Y_t = \mathbf{1}\{H_t > 0\}$, Proposition A.18 gives $\mathbb{E}_P[Y_1Y_2] - \mathbb{E}_Q[Y_1Y_2] = (\arcsin(1/4) - \arcsin(1/5))/(2\pi) > 0$.

We can therefore state the main separation theorem.

Theorem A.19 (Temporal irreducibility of the hard-spiking kernel)**.** *There exist two spike-path distributions $P, Q \in \mathcal{P}(\{0,1\}^2)$ such that $P \equiv_{\mathrm{rate}} Q$, but a single infinite-width Gaussian hard-spiking layer produces distinct temporal kernels: $\mathbf{Q}_{\mathrm{out}}^P \neq \mathbf{Q}_{\mathrm{out}}^Q$.*

More quantitatively, $\|\mathbf{Q}_{\mathrm{out}}^P - \mathbf{Q}_{\mathrm{out}}^Q\|_{\mathrm{op}} \geq (\arcsin(1/4) - \arcsin(1/5))/(2\pi) > 0$. Hence the hard-spiking temporal kernel map is not rate reducible.

This result is stronger than showing that two spike trajectories with identical mean rates can be distinguished. The complete spike-count distribution is identical under P and Q. The only distinction lies in *where in time* the spikes occur.

A.3.5 TEMPORAL INFORMATION CAN ALTER THE DOWNSTREAM RATE KERNEL

The same $T = 2$ construction remains separated after an output rate readout. Define $\overline{Y} = (Y_1 + Y_2)/2$. Since $\mathbb{E}[Y_1] = \mathbb{E}[Y_2] = 1/2$, $\mathbb{E}[\overline{Y}^2] = \frac{1}{4}\mathbb{E}[Y_1 + Y_2 + 2Y_1Y_2] = \frac{3}{8} + \frac{1}{4\pi}\arcsin(\rho)$. Therefore

$$\mathbb{E}_P[\overline{Y}^2] - \mathbb{E}_Q[\overline{Y}^2] = \frac{\arcsin(1/4) - \arcsin(1/5)}{4\pi} > 0. \tag{56}$$

For a Gaussian linear readout $f(x) = \sigma_{\text{out}} n^{-1/2} \sum_i a_i \overline{Y}_i(x)$, $a_i \sim \mathcal{N}(0,1)$, its self-kernel satisfies

$$K^P_{\text{rate,out}} - K^Q_{\text{rate,out}} = \sigma^2_{\text{out}} \frac{\arcsin(1/4) - \arcsin(1/5)}{4\pi} > 0. \tag{57}$$

Corollary A.20 (Temporal-to-rate conversion)**.** *Two input spike populations can have identical complete firing-rate distributions while inducing different downstream firing-rate kernels after one hard-spiking wide layer.*

Thus the hard threshold converts input structure orthogonal to rate into a difference observable by a rate readout.

A.3.6 Strengthening the equivalence: equal per-bin rates

The construction of Section A.3.4 equalises the spike-count law, but its first temporal moments differ, $\mathbf{m}_P = (0.8, 0.2)^\top$ against $\mathbf{m}_Q = (0.5, 0.5)^\top$. Since $\mathbf{m}$ is the vector of *per-bin firing rates*, a rate model that resolves time at all separates the two populations already at first order, and the separation of Theorem A.19 cannot be attributed to second-order temporal structure alone. We therefore strengthen the equivalence.

Definition A.21 (Strong rate equivalence)**.** Two spike-path distributions are *strongly rate equivalent* if they share both the spike-count law and the per-bin marginals, $\mathbf{m} = \widetilde{\mathbf{m}}$.

When the shared per-bin marginals are all equal to a common $p > 0$, the uncentered correlations $\rho_{tt'} = Q_{tt'}/p$ lie in $[0,1]$, and their sum is pinned, $\sum_{t \neq t'} \rho_{tt'} = \frac{\mathbb{E}[C^2] - \mathbb{E}[C]}{p}$, where C is the spike count. The diagonal of the Gaussian drive is fixed too, so both populations produce $\mathbb{E}[Y_t] = \frac{1}{2}$ at a zero-threshold layer. Only the *spread* of the correlation spectrum can differ. That the second-order statistic is the irreducible one has an exact counterpart in the mean-field description of spiking populations, whose macroscopic dynamics close only when a synchrony variable is carried alongside the rate (Montbrió et al., 2015).

A first observation is that the horizon must be long enough.

Proposition A.22 (No strong separation at $T = 2$)**.** *For $T = 2$ the constraints of Definition A.21 determine the joint law completely: no two distinct strongly rate-equivalent distributions on $\{0,1\}^2$ exist with different second moments.*

Proof. Write $p_{ab} = \mathbb{P}[\mathbf{s} = (a,b)]$ for $a, b \in \{0,1\}$, so the law has four atoms summing to one. The count law fixes $p_{00} = \mathbb{P}[C = 0]$ and $p_{11} = \mathbb{P}[C = 2]$ outright, and fixes $p_{01} + p_{10} = \mathbb{P}[C = 1]$. The per-bin marginals give $m_1 = p_{10} + p_{11}$ and $m_2 = p_{01} + p_{11}$, so $p_{10} = m_1 - p_{11}$ and $p_{01} = m_2 - p_{11}$ are determined as well. All four atoms are therefore pinned by the count law together with the marginals, and two strongly rate equivalent laws on $\{0,1\}^2$ coincide. In particular $Q_{12} = p_{11}$ agrees, so no second-order temporal difference survives at $T = 2$. □

This explains why the construction of Section A.3.4 was forced to break the marginals: at $T = 2$ it had no alternative. We use a particularly symmetric strong separation at $T = 4$.

Proposition A.23 (Equal-marginal separating pair)**.** *Let $T = 4$ and let every sample carry exactly two spikes, so the count law is deterministic and $m_t = \frac{1}{2}$ for both members. Let P be uniform over all six weight-two vectors, and $\widetilde{P}$ uniform over the four-cycle $\{12, 23, 34, 14\}$. Then P and $\widetilde{P}$ are strongly rate equivalent with the same mean correlation $\bar{\rho} = \frac{1}{3}$, while $\rho^P_{tt'} \equiv \frac{1}{3}$ and $\rho^{\widetilde{P}}_{tt'}$ equals $\frac{1}{2}$ on the cycle edges and 0 otherwise.*

Proof. Both laws are supported on weight-two vectors, so $C \equiv 2$ and the count law is deterministic and common. For P, each index t lies in exactly three of the six weight-two vectors, so $m_t = 3/6 = \frac{1}{2}$, and each unordered pair $\{t, t'\}$ is the support of exactly one of them, so $Q_{tt'} = 1/6$ and $\rho^P_{tt'} = (1/6)/\frac{1}{2} = \frac{1}{3}$. For $\widetilde{P}$, each index lies in exactly two of the four cycle edges, so again $m_t = 2/4 = \frac{1}{2}$; a pair $\{t, t'\}$ that is a cycle edge is the support of exactly one of them, giving $Q_{tt'} = 1/4$ and $\rho^{\widetilde{P}}_{tt'} = \frac{1}{2}$, and a pair that is a diagonal is the support of none, giving 0. Averaging over the $T(T-1) = 12$ ordered off-diagonal pairs, $\bar{\rho}^P = \frac{1}{3}$ and $\bar{\rho}^{\widetilde{P}} = (8 \cdot \frac{1}{2} + 4 \cdot 0)/12 = \frac{1}{3}$. □

Because the two members differ only in the spread of a pinned correlation spectrum, and $\arcsin$ is strictly convex on $[0, 1]$, the direction of the resulting separation is forced.

Theorem A.24 (Quantitative equal-marginal separation)**.** *For strongly rate-equivalent $P, \widetilde{P}$ with positive per-bin marginals and a centered, memoryless zero-threshold hard-spiking layer with $\mathbf{B} = \mathbf{I}$ and $\sigma > 0$, put $\rho_{tt'} = Q_{tt'}/\sqrt{m_t m_{t'}}$. The unit-vector rate contractions differ by*

$$\mathbf{v}_0^\top \left(\mathbf{Q}_{\text{out}} - \widetilde{\mathbf{Q}}_{\text{out}}\right) \mathbf{v}_0 = \frac{1}{2\pi T} \sum_{t \neq t'} \left[\arcsin \rho_{tt'} - \arcsin \widetilde{\rho}_{tt'}\right], \tag{58}$$

which is strictly positive whenever the correlation spectrum of P is a strict mean-preserving spread of that of $\widetilde{P}$. For comparison with a constant-correlation baseline $\bar{\rho} \in (0, 1)$, write $\rho_{tt'} = \bar{\rho} + \delta_{tt'}$ with mean-zero perturbations and $\max |\delta_{tt'}| \to 0$. Then

$$\mathbf{v}_0^\top (\mathbf{Q}_{\text{out}} - \mathbf{Q}_{\text{out,baseline}}) \mathbf{v}_0 = \frac{T-1}{4\pi} \frac{\bar{\rho}}{(1-\bar{\rho}^2)^{3/2}} \operatorname{Var}(\delta) + O\left(\frac{1}{T} \sum_{t \neq t'} |\delta_{tt'}|^3\right). \tag{59}$$

This expansion uses a fixed interior baseline and bounded third derivative in its neighborhood. It is not a formula for arbitrary pairs from their single variance alone.

Proof. Write $\mathbf{Y} = \Psi_{\text{LIF}}(\mathbf{Z})$ for the output of the memoryless zero-threshold layer, so $Y_t = \mathbf{1}\{Z_t \geq 0\}$ with $\mathbf{Z}$ centered Gaussian with correlation $\rho_{tt'}$. Expanding the contraction, $\mathbf{v}_0^\top \mathbf{Q}_{\text{out}} \mathbf{v}_0 = \frac{1}{T}[\sum_t Q_{\text{out}}(t,t) + \sum_{t \neq t'} Q_{\text{out}}(t,t')]$. The diagonal entries are $\mathbb{E}[Y_t^2] = \mathbb{E}[Y_t] = \mathbb{P}[Z_t \geq 0] = \frac{1}{2}$ for every t and for both members, because the coordinates are centered with positive variance; they cancel in the difference. The off-diagonal entries are supplied by Proposition A.18, $Q_{\text{out}}(t,t') = \frac{1}{4} + \arcsin(\rho_{tt'})/(2\pi)$, whose constant $\frac{1}{4}$ also cancels. Subtracting gives (58). For the sign, a mean-preserving spread leaves $\sum_{t \neq t'} \rho_{tt'}$ unchanged, and $\arcsin$ is strictly convex on $[0, 1]$, so a strict mean-preserving spread strictly increases $\sum_{t \neq t'} \arcsin \rho_{tt'}$ by Jensen's inequality.

For (59), let $\mathbf{Q}_{\text{out,baseline}}$ denote the output second moment generated by the constant-correlation matrix $\rho_{tt'} \equiv \bar{\rho}$ $(t \neq t')$, which by the previous paragraph has the same diagonal. On a compact neighbourhood of $\bar{\rho}$ in $(-1, 1)$, $\arcsin$ is smooth with $\arcsin''(x) = x(1-x^2)^{-3/2}$, so $\arcsin(\bar{\rho} + \delta) = \arcsin \bar{\rho} + \frac{\delta}{\sqrt{1-\bar{\rho}^2}} + \frac{\bar{\rho}}{2(1-\bar{\rho}^2)^{3/2}} \delta^2 + O(|\delta|^3)$, with the remainder constant the supremum of $|\arcsin'''|$ over that neighbourhood. Summing over the $T(T-1)$ ordered off-diagonal pairs, the first-order term vanishes because the $\delta_{tt'}$ have mean zero, and $\sum_{t \neq t'} \delta_{tt'}^2 = T(T-1) \operatorname{Var}(\delta)$ by definition of the variance over those pairs. Dividing by $2\pi T$ as in (58) gives the coefficient $(T-1)\bar{\rho}/\left(4\pi(1-\bar{\rho}^2)^{3/2}\right)$ and the stated remainder. The expansion is asymptotic as $\max |\delta_{tt'}| \to 0$ at fixed $\bar{\rho}$; it is not a formula for a pair at fixed separation: for the pair of Proposition A.23, with P as the baseline $\bar{\rho} = \frac{1}{3}$ and $\operatorname{Var}(\delta) = \frac{1}{18}$ for $\widetilde{P}$, its leading term is 5.275×10^{-3} against the exact 4.406×10^{-3}. □

Unlike Theorem A.19, this statement supplies a magnitude and a sign rather than an inequality, and both are attributable to second-order temporal structure alone. For the pair of Proposition A.23, the cycle law has the larger unit-vector rate contraction, by approximately 4.406×10^{-3}; see (8) for the direction and formula. The actual firing-rate kernel includes a further factor $1/T$. The cycle second moment has spectrum $\{0, 1/2, 1/2, 1\}$ and is singular, but at the memoryless point every coordinate threshold has positive variance. Its boundary-nullness therefore holds. A nonzero gap certifies separation; its magnitude need not imply a substantial predictive gain.

A.3.7 Persistence for genuine hard LIF dynamics

The explicit separation above used the memoryless point $\lambda = \rho = \vartheta = 0$ to obtain a closed-form expression. We next show that temporal irreducibility is not confined to this isolated parameter choice. Let $\boldsymbol{\eta} = (\lambda, \rho, \vartheta)$ denote the LIF parameter vector, and let $\mathbf{Q}_{\text{out}}^P(\boldsymbol{\eta})$, $\mathbf{Q}_{\text{out}}^Q(\boldsymbol{\eta})$ be the output temporal kernels generated from the two Gaussian input laws in (54)–(55).

Define the kernel separation $\Delta(\boldsymbol{\eta}) = \|\mathbf{Q}_{\text{out}}^P(\boldsymbol{\eta}) - \mathbf{Q}_{\text{out}}^Q(\boldsymbol{\eta})\|_{\text{op}}$.

Lemma A.25 (Continuity in LIF parameters)**.** *Suppose the Gaussian input fields are nondegenerate. Then $\boldsymbol{\eta} \mapsto \mathbf{Q}^{P}_{\mathrm{out}}(\boldsymbol{\eta})$ and $\boldsymbol{\eta} \mapsto \mathbf{Q}^{Q}_{\mathrm{out}}(\boldsymbol{\eta})$ are continuous at every parameter point for which the associated threshold boundaries have zero Gaussian measure. Consequently, $\boldsymbol{\eta} \mapsto \Delta(\boldsymbol{\eta})$ is continuous.*

Proof. For every fixed output spike pattern $\mathbf{s}$, Lemma A.2 represents the corresponding LIF region as $\mathcal{R}_{\mathbf{s}}(\boldsymbol{\eta}) = \{\mathbf{h} : \mathbf{C}_{\mathbf{s}}(\boldsymbol{\eta})\mathbf{h} \le \mathbf{d}_{\mathbf{s}}(\boldsymbol{\eta})\}$, up to boundary conventions, where $\mathbf{C}_{\mathbf{s}}(\boldsymbol{\eta})$ and $\mathbf{d}_{\mathbf{s}}(\boldsymbol{\eta})$ depend continuously on $\boldsymbol{\eta}$.

For a nondegenerate Gaussian field, every affine hyperplane has probability zero. Small perturbations of $\boldsymbol{\eta}$ therefore perturb the polyhedral regions without creating positive probability mass on their boundaries. Dominated convergence, equivalently continuity of Gaussian probabilities of moving polyhedra at boundary-null configurations, yields $\mathbb{P}[\mathbf{H} \in \mathcal{R}_{\mathbf{s}}(\boldsymbol{\eta}_n)] \to \mathbb{P}[\mathbf{H} \in \mathcal{R}_{\mathbf{s}}(\boldsymbol{\eta})]$ whenever $\boldsymbol{\eta}_n \to \boldsymbol{\eta}$.

Since the spike-pattern space is finite, all first and second output moments are continuous. □

Combining Lemma A.25 with Theorem A.19 gives:

Corollary A.26 (Open-set temporal irreducibility)**.** *There exists an open neighborhood $\mathcal{U} \subset \mathbb{R}^3$ of $(\lambda, \rho, \vartheta) = (0, 0, 0)$ such that $\Delta(\lambda, \rho, \vartheta) > 0$ for every $(\lambda, \rho, \vartheta) \in \mathcal{U}$. In particular, temporal irreducibility persists for hard LIF neurons with strictly positive leak and reset parameters.*

Proof. Theorem A.19 gives $\Delta(0, 0, 0) > 0$. The fields (54)–(55) are nondegenerate, since $\det \mathbf{Q}_P = 0.15$ and $\det \mathbf{Q}_Q = 0.24$ by (52) and $\sigma > 0$ by (53). By Lemma A.25, Δ is continuous at the origin. Hence there exists an open neighborhood $\mathcal{U}$ in which $\Delta(\boldsymbol{\eta}) > \frac{1}{2}\Delta(\mathbf{0}) > 0$. □

Therefore, temporal irreducibility holds on an open neighbourhood of hard-LIF parameters and is not confined to the memoryless limit.

How large is that open set? Corollary A.26 is proved by continuity and therefore says nothing about the size of the neighbourhood. Measuring it with the strong equal-marginal pair of Proposition A.23, so that whatever survives is second-order temporal structure alone, the answer is that the neighbourhood is the whole tested cube (Figure 13).

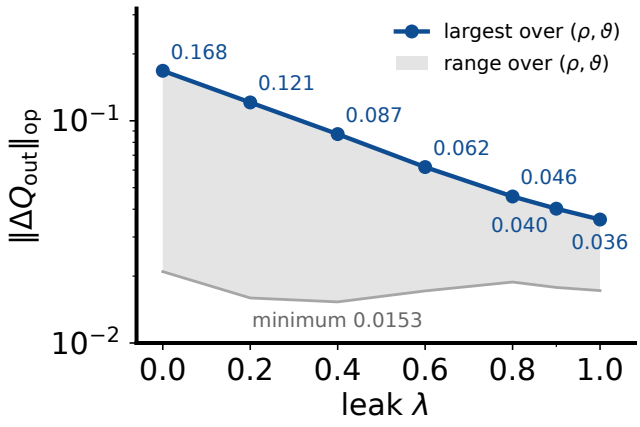


Figure 13: Separation $\|\Delta\mathbf{Q}_{\mathrm{out}}\|_{\mathrm{op}}$ of the equal-marginal pair against the leak, over (ρ, ϑ).

Across the 224 points $\lambda \in \{0, 0.2, 0.4, 0.6, 0.8, 0.9, 1\}$, $\rho \in \{0, 0.25, 0.5, 1\}$ and $\vartheta \in \{-0.4, -0.2, \ldots, 1.0\}$, with $\mathbf{B} = \mathbf{I}$, $\mu = 0$, $\mathbf{b} = \mathbf{0}$, $\sigma = 1$ and initial membrane value 0, and the separation estimated on 2^{17} quasi-random samples of a single scrambling shared by the two members and by all points, each member's field formed from them through the Cholesky factor of its covariance plus $10^{-6}\mathbf{I}$, every point separates by more than 10^{-2}, with minimum 0.0153 and maximum 0.168. The maximum separation decreases monotonically in the leak but never vanishes, including at the integrator endpoint: the exact Gaussian-orthant value, evaluated at the 119 grid points with $\rho \in \{0, 0.5\}$ or $\rho = \vartheta = 1$, is strictly positive at each of them, with minimum

1.71×10^{-2} on the face $\lambda = 1$, and the soft-reset points $\lambda = 1$, $\rho = \vartheta \in \{0.5, 1\}$ give 3.03×10^{-2} and 2.16×10^{-2}.

Pointwise the decrease in the leak fails in five of the thirty-two (ρ, ϑ) columns, all at high threshold, and the failure includes the column carrying the reported minimum, where the exact values run 0.02094, 0.01579, 0.01531, 0.01716, 0.01938, 0.02073, 0.02160 across λ and rise 41% from $\lambda = 0.4$ onward. In the threshold the separation is not monotone: at $\lambda = \rho = 0$ the map $\mathbf{h} \mapsto -\mathbf{h}$ sends the firing pattern at ϑ to the complementary pattern at $-\vartheta$, and equal marginals cancel the first-moment term, so the separation is an *even* function of ϑ (verified at the cell-probability level for both members at $\vartheta = 0.2$ and 0.4: each firing-pattern probability at ϑ matches that of the complementary pattern at $-\vartheta$ to within 8×10^{-8}, with each cell probability evaluated once by randomised lattice quadrature at tolerance 10^{-7}). Its exact values 0.08773, 0.10576, 0.11258, 0.10576, 0.08773 at $\vartheta = -0.4, \ldots, 0.4$ have an interior maximum at balanced firing, rising 28% from the left edge. So the maximum separation decreases monotonically in the leak, while at $\lambda = \rho = 0$ it falls away in the threshold on either side of balanced firing.

Outside the reported cube the decay continues without ever reaching zero: at $\vartheta = 3, 3.5, 4$ the exact values are 2.1×10^{-7}, 2.2×10^{-9} and 1.2×10^{-11}, where a Monte-Carlo estimate on 2^{17} quasi-random samples of another single scrambling, shared by the two members, instead pins to multiples of 2^{-17} and reports 7.6×10^{-6} and then exactly zero. These points lie outside the range Figure 13 covers and are recorded only to show that the positivity is not an artifact of where the grid was cut.

Together with the numerical depth diagnostics of Section A.2.10, the tested configurations show that *a single layer separates at every tested point of the cube, and depth contracts at every tested point of the cube.* These are observations over the tested configurations, not universal parameter or depth guarantees.

A.3.8 A GENERAL TEMPORAL WITNESS CRITERION

The explicit construction can be abstracted into a more general sufficient condition.

Let two spike populations have descriptors $c = (\mathbf{m}, \mathbf{Q})$, $\widetilde{c} = (\widetilde{\mathbf{m}}, \widetilde{\mathbf{Q}})$, and suppose they are rate equivalent at the level relevant to a downstream rate representation: $\mathbf{\Pi m} = \mathbf{\Pi}\widetilde{\mathbf{m}}$, $\mathbf{\Pi Q \Pi} = \mathbf{\Pi}\widetilde{\mathbf{Q}}\mathbf{\Pi}$.

Consider a zero-mean Gaussian mixing layer, $\mu = 0$, with temporal operator $\mathbf{B}$. Its two preactivation covariances are $\mathbf{\Sigma} = \sigma^2 \mathbf{B Q B}^\top$, and $\widetilde{\mathbf{\Sigma}} = \sigma^2 \mathbf{B}\widetilde{\mathbf{Q}}\mathbf{B}^\top$.

Corollary A.27 (Architecture-filtered temporal witness)**.** *Suppose there exist two downstream temporal coordinates $r \neq q$ such that all corresponding variances are strictly positive. Define $\rho_{rq} := \Sigma_{rq}/\sqrt{\Sigma_{rr}\Sigma_{qq}}$ and $\widetilde{\rho}_{rq} := \widetilde{\Sigma}_{rq}/\sqrt{\widetilde{\Sigma}_{rr}\widetilde{\Sigma}_{qq}}$, and suppose $\rho_{rq} \neq \widetilde{\rho}_{rq}$. Then a centered memoryless zero-threshold hard-spiking layer has $\mathbf{Q}_{\mathrm{out}} \neq \widetilde{\mathbf{Q}}_{\mathrm{out}}$, and the operator-norm bound in* (51) *holds with these filtered correlations.*

Proof. Apply Proposition A.18 to the Gaussian covariances $\mathbf{\Sigma} = \sigma^2 \mathbf{B Q B}^\top$ and $\widetilde{\mathbf{\Sigma}} = \sigma^2 \mathbf{B}\widetilde{\mathbf{Q}}\mathbf{B}^\top$. □

Corollary A.27 identifies the exact role of the temporal operator. It is not sufficient for the input populations merely to have different non-DC covariance. Their difference must survive the transformation $\mathbf{Q} \mapsto \mathbf{B Q B}^\top$. Thus the relevant architecture-dependent temporal discrepancy is $\mathbf{B}(\mathbf{Q} - \widetilde{\mathbf{Q}})\mathbf{B}^\top$.

If this matrix alters a normalized correlation seen by a hard threshold, temporal information becomes functionally observable. This complements the leakage quantity of Section A.2.4: $\mathcal{E}_{\mathrm{temp}}$ controls when discarded temporal information is guaranteed to be small, whereas Corollary A.27 provides a positive witness when surviving temporal information is guaranteed to matter.

A.3.9 FROM TEMPORAL MODES TO FUNCTION-SPACE SEPARATION

For a temporal readout $\mathbf{a}$, the induced kernel is $K_{\mathbf{a}}(x, x') = \sigma_{\mathrm{out}}^2 \mathbf{a}^\top \mathbf{Q}^L(x, x')\mathbf{a}$. Hence any symmetric temporal-kernel difference with a nonzero quadratic direction is detected by some readout.

In the explicit construction, even the rate readout $\mathbf{a}_{\mathrm{rate}} = (1,1)^\top/2$ detects the difference by (57); rate-equivalent input populations therefore induce distinct Gaussian-process kernels.

A.3.10 From kernel gap to information the readout can use

A kernel difference need not affect a target if the readout suppresses it. To obtain a nested comparison, let $\mathbf{\Lambda} \succeq 0$ weight temporal readouts and define $K_{\mathbf{\Lambda}}(x,x') = \mathrm{tr}(\mathbf{\Lambda}\mathbf{Q}(x,x'))$. Rank-one $\mathbf{\Lambda} = \mathbf{a}\mathbf{a}^\top$ recovers a fixed readout; $\mathbf{P}_0$, $\mathbf{P}_\Delta$ and $\mathbf{I}$ give rate, finite-resolution and full temporal readouts.

Proposition A.28 (Monotone readout lattice)**.** *Each $K_{\mathbf{\Lambda}}$ is positive semidefinite, and $\mathbf{\Lambda}_1 \succeq \mathbf{\Lambda}_2$ implies $K_{\mathbf{\Lambda}_1} - K_{\mathbf{\Lambda}_2} = K_{\mathbf{\Lambda}_1 - \mathbf{\Lambda}_2} \succeq 0$. Thus $K_{\mathbf{I}} = K_{\mathbf{P}_0} + K_{\mathbf{P}_\perp}$, and $\Delta' \mid \Delta$ implies $K_{\mathbf{P}_{\Delta'}} \succeq K_{\mathbf{P}_\Delta}$ and $\mathcal{H}_{\mathbf{P}_\Delta} \subseteq \mathcal{H}_{\mathbf{P}_{\Delta'}}$.*

Proof. Write $\mathbf{\Lambda} = \sum_k \mathbf{a}_k\mathbf{a}_k^\top$. Each summand is the kernel of the scalar spike feature $\mathbf{a}_k^\top\mathbf{S}(\cdot)$; positivity, linearity and the nesting of block subspaces give the claims. □

We use standard kernel and spectral learning notation (Cristianini et al., 2001; Caponnetto & De Vito, 2007; Canatar et al., 2021). Let $T_{\mathbf{\Lambda}}$ be the integral operator of $K_{\mathbf{\Lambda}}$ on $L^2(\rho_X)$, and let $V_{\mathbf{\Lambda}} = \overline{\mathrm{range}\, T_{\mathbf{\Lambda}}}$. For a target f, denote the risk of a kernel-ridge predictor $\widehat{f} \in \mathcal{H}_{\mathbf{\Lambda}}$ by $R(f; K_{\mathbf{\Lambda}}) = \|\widehat{f} - f\|^2_{L^2(\rho_X)}$.

Theorem A.29 (Deficiency lower bound)**.** *For every sample size, ridge parameter and training set, $R(f; K_{\mathbf{\Lambda}}) \geq \|(\mathbf{I} - \mathbf{P}_{V_{\mathbf{\Lambda}}})f\|^2_{L^2(\rho_X)}$.*

Proof. Every kernel-ridge predictor lies in $\mathcal{H}_{\mathbf{\Lambda}} \subseteq V_{\mathbf{\Lambda}}$; orthogonal projection gives the bound. □

Corollary A.30 (Rate-deficient encoders)**.** *If the rate quotient and bias paths are constant across inputs, then the rate-projected recursion gives a constant surrogate kernel, $V_{\mathrm{rate\text{-}proj}}$ contains only constants, and*

$$\inf_N R(f; K_{\mathrm{rate\text{-}proj}}) \geq \mathrm{Var}_{\rho_X}(f). \tag{60}$$

The full temporal kernel can still separate inputs whenever Corollary A.27 applies.

Proof. Proposition A.7 makes every projected layer depend only on the common rate quotient, so induction gives $K_{\mathrm{rate\text{-}proj}}(x,x') \equiv \kappa$. Its feature space is the span of constants when $\kappa > 0$, and the zero space otherwise; Theorem A.29 yields (60). □

This is an observability bound, not an unconditional ordering of finite-sample risks. A general comparison would additionally require assumptions on the target, noise, regularization and the two kernels' eigenfunctions.

A.3.11 Mode conversion and the final criterion

A nonlinear spike map can convert nonconstant temporal structure into a downstream rate component: $\mathbf{P}_0\mathcal{K}_{\mathrm{spike}}(\mathbf{Q}_{\mathrm{dc}} + \mathbf{Q}_{\mathrm{temp}})\mathbf{P}_0 \neq \mathbf{P}_0\mathcal{K}_{\mathrm{spike}}(\mathbf{Q}_{\mathrm{dc}})\mathbf{P}_0$. Equation (56) gives an explicit instance. Together, the collapse and witness results identify the full path $\mathbf{Q}_{\mathrm{temp}} \xrightarrow{\mathbf{B}} \mathbf{\Sigma}_{\mathrm{temp}} \xrightarrow{\Psi_{\mathrm{LIF}}} \mathbf{Q}_{\mathrm{out}} \xrightarrow{\text{readout}} K_{\mathrm{SNN}}$. An SNN is rate sufficient when this path factors through the rate quotient, exactly or approximately. Timing is irreducible when a nonconstant component survives every stage and changes the readout kernel. The distinction is therefore not the presence of binary spikes, but whether temporal information survives the architecture and remains observable at its output.

A.4 Extended Experimental Protocol and Diagnostics

This section gives the implementation details, controls and measured outcomes for the tests summarized in Table 3. The first experiments test the temporal limit, rate collapse and temporal separation; the later experiments ask whether the same quantities explain finite trained SNNs. Each result states its sample or seed count and any numerical floor or failed prediction.

Panel index. Main-text panels map to their detailed checks as follows: Figure 2(a,b,c) to Appendices A.4.2 and A.4.11; Figure 3(a,b,c) to Appendices A.2.10, A.4.3 and A.5; Figure 4(a,b,c) to Appendices A.2.12, A.3.2 and A.3.7, with the memoryless value in Appendix A.4.4; Figure 5(a,b,c) to Appendices A.4.6, A.4.5 and A.4.10; and Figure 6(a,b) with Table 1 to Appendices A.4.8 and A.4.9. The remaining checkpoint figures and Table 2 are documented in Appendix A.4.9.

A.4.1 Empirical temporal kernels and diagnostics

For a network of width n_ℓ, we estimate the layer-ℓ temporal spike kernel and its M-input block form as $\widehat{\mathbf{K}}^\ell(x,x') = n_\ell^{-1}\sum_{i=1}^{n_\ell}\mathbf{S}_i^\ell(x)\mathbf{S}_i^\ell(x')^\top \in \mathbb{R}^{T\times T}$ and $\widehat{\mathbf{Q}}^\ell = n_\ell^{-1}\sum_{i=1}^{n_\ell}\mathbf{S}_i^\ell(\mathbf{S}_i^\ell)^\top$.

All diagnostics below are computed directly from spike trajectories and therefore require neither gradients nor surrogate activations.

Temporal kernel mass. We quantify how much of a representation lies outside the rate/DC subspace by $\widehat{\mathfrak{M}}^\ell_{\mathrm{temp}} = \|\widehat{\mathbf{Q}}^\ell - \mathbf{\Pi}\widehat{\mathbf{Q}}^\ell\mathbf{\Pi}\|_F/\|\widehat{\mathbf{Q}}^\ell\|_F$.

A value close to zero indicates an approximately rate-like representation, whereas a large value indicates substantial temporal structure.

For approximately stationary trajectories, we additionally report the spectral version of the trace-normalized mass (45) $\widehat{\mathfrak{M}}_{\mathrm{temp,spec}} = \sum_{\omega\neq 0}\widehat{\lambda}_\omega/\sum_\omega\widehat{\lambda}_\omega$, where $\{\widehat{\lambda}_\omega\}$ are the temporal Fourier eigenvalues of the empirical kernel.

Transmitted temporal energy. Theorem A.11 identifies the temporal information that actually reaches the next layer. We estimate $\widehat{\mathcal{E}}^\ell_{\mathrm{temp}}$ by substituting the empirical descriptor $(\widehat{\mathbf{m}}^\ell, \widehat{\mathbf{Q}}^\ell)$ into (36). Because its absolute scale depends on the layer parameterization, we report the normalized quantity $\widehat{\eta}^\ell_{\mathrm{temp}} = \widehat{\mathcal{E}}^\ell_{\mathrm{temp}}/(\mu_{\ell+1}^2\|\mathbf{B}_\ell\widehat{\mathbf{m}}^\ell\|_2^2 + \sigma_{\ell+1}^2\,\mathrm{tr}(\mathbf{B}_\ell\widehat{\mathbf{Q}}^\ell\mathbf{B}_\ell^\top) + \epsilon)$. Here, and in (61) below, $\epsilon > 0$ is a small constant that keeps the ratio defined when the rest of the denominator vanishes.

The distinction between $\widehat{\mathfrak{M}}_{\mathrm{temp}}$ and $\widehat{\eta}_{\mathrm{temp}}$ is important. The former measures how much temporal structure is present; the latter measures how much of that structure is transmitted by the architecture.

Rate-collapse gap. For each network we use the rate-projected surrogate of Section A.2.3, keeping all non-operator parameters unchanged. Let $\widehat{\mathbf{K}}_{\mathrm{full}}$ and $\widehat{\mathbf{K}}_{\mathrm{dc}}$ denote the empirical readout Gram matrices of the original and projected networks. We define $\widehat{\Delta}_{\mathrm{collapse}} = \|\widehat{\mathbf{K}}_{\mathrm{full}} - \widehat{\mathbf{K}}_{\mathrm{dc}}\|_F/\|\widehat{\mathbf{K}}_{\mathrm{full}}\|_F$.

When $\|\widehat{\mathbf{K}}_{\mathrm{full}}\|_F$ has a uniform positive lower bound, Theorem A.12 implies that this relative gap vanishes with the transmitted temporal energy; without that condition, it remains a diagnostic rather than a theorem-level limit.

Temporal-to-rate conversion. Temporal structure need not remain temporal. Section A.3.11 shows that hard spiking can convert a non-DC input difference into a downstream DC difference.

For an intervention producing two spike populations P and Q with identical rate statistics, we therefore measure

$$\widehat{\mathcal{C}}_{\mathrm{T}\to\mathrm{R}} = \frac{\|\mathbf{\Pi}(\widehat{\mathbf{Q}}^P_{\mathrm{out}} - \widehat{\mathbf{Q}}^Q_{\mathrm{out}})\mathbf{\Pi}\|_F}{\|\widehat{\mathbf{Q}}^P_{\mathrm{out}}\|_F + \epsilon}. \tag{61}$$

A positive value directly measures conversion of purely temporal input information into downstream rate-level information.

A.4.2 Experiment I: finite-width convergence to the temporal NNGP

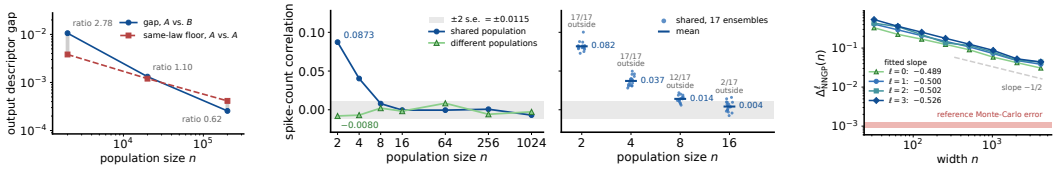
Setup. We compare finite hard-LIF networks with the recursion of Appendix A.1. The primary run uses depth three, $T = 4$, $M = 6$, $(\lambda, \rho, \vartheta) = (0.5, 0.3, 0.35)$, identity temporal operators, Gaussian weights with $(\mu, \sigma) = (0, 1)$, and eight independent networks per width. The recursion applies the exact hard-LIF map to Gaussian drives; no smooth threshold is used. At layer ℓ we report

$$\Delta^\ell_{\mathrm{NNGP}}(n) = \frac{\|\widehat{\mathbf{Q}}^\ell_n - \mathbf{Q}^\ell_\infty\|_F}{\|\mathbf{Q}^\ell_\infty\|_F}. \tag{62}$$

Finite-width recursion. For widths $32, 64, \ldots, 4096$, the log–log slopes of (62) are $-0.489, -0.500, -0.502, -0.526$ at layers zero to three. The same $n^{-1/2}$ slope occurs at layer zero, where no network recursion is present, so the exponent is a sampling baseline rather than independent evidence for the recursion. A power test therefore perturbs the predicted off-diagonal entries by a known relative error δ: the smallest detectable δ falls from above 0.20 at width 512 to 0.05 at 8192, while the unperturbed error falls from 0.173 at width 32 to 0.009 at 8192 (Figure 2b). The recursion's two-scramble numerical floor is about 1.2×10^{-3} per layer.

Sufficiency and propagation of chaos. To test dependence on only $(\mathbf{m}, \mathbf{Q})$, we compare the uniform laws on even- and odd-parity spike vectors at $T = 3$. They have identical first and second moments but third moments 0 and $1/4$. After one exact layer, their descriptor gap remains within 0.62–2.78 times the same-law sampling floor, and exact cell probabilities differ by at most 9.8×10^{-8}, the 10^{-7} quadrature tolerance. In a separate $T = 4$ Bernoulli experiment over 3×10^4 populations, the spike-count correlation of two output neurons falls from 0.0873 at population size two to the null band by size eight; across 17 ensembles it averages 0.014 at size eight and 0.004 at 16.

Smooth-threshold control. Replacing the hard threshold by a logistic function with slope β does not reproduce the same kernel unless the surrogate is nearly discontinuous. Relative errors at layers one to three are $(0.21, 0.28, 0.34)$ for $\beta = 1$, $(0.055, 0.085, 0.106)$ for $\beta = 10$, and $(0.0007, 0.0011, 0.0012)$ for $\beta = 1000$.



(a) equal moments (b) one ensemble (c) 17 ensembles (d) kernel convergence

Figure 14: **Finite-width checks.** (a) Equal-moment populations versus the same-law floor; (b, c) two-neuron dependence; (d) kernel convergence with width.

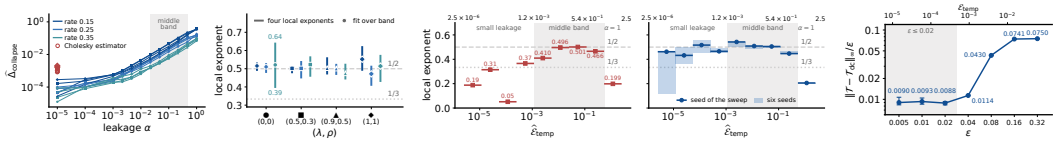
A.4.3 Experiment II: controlled rate collapse

Operator sweep. We suppress nonconstant temporal directions with $\mathbf{A}_\alpha = \mathbf{A}\mathbf{P}_0 + \alpha\mathbf{A}\mathbf{P}_\perp$, $\alpha \in [0, 1]$. Thus $\alpha = 0$ satisfies exact rate factorization and $\widehat{\mathcal{E}}_{\mathrm{temp}} = O(\alpha^2)$. The primary sweep uses depth three, $T = M = 4$, $(\lambda, \rho) = (0.5, 0.3)$, $(\mu, \sigma) = (0, 1)$, a causal exponential base operator with time constant three, ten logarithmically spaced nonzero values of α, and layerwise threshold calibration to firing rate 0.25.

Collapse and numerical floor. The relative output-descriptor gap is exactly zero at $\alpha = 0$ and reaches 0.147 at $\alpha = 1$. The original Cholesky coupling plateaus near 10^{-3} as $\alpha \to 0$; increasing samples reduces this plateau, identifying it as estimator noise. Coupling the two arms through the common square-root construction of (34)–(35) and tightening threshold calibration makes the six-seed mean nondecreasing from $\alpha = 0$ in all 12 combinations of four membrane/reset settings and three firing rates. In 68 of 72 runs the gap is numerically zero at $\alpha = 10^{-7}$, and the other four are below 1.1×10^{-4}.

Scaling and theorem-level bound. Below saturation, local slopes of the collapse gap against transmitted energy lie between 0.39 and 0.64, with fitted slopes 0.47–0.55. This is empirical local square-root scaling, not a universal exponent. A separate one-layer test uses $\mathbf{B} = \mathbf{\Pi} + \varepsilon\mathbf{\Pi}_\perp$, $T = 4$, $M = 3$, and 2.56×10^8 shared samples. The bound of Theorem A.11 holds at every tested ε, with slack 397–1.3×10^4. For $\varepsilon \le 0.02$, the measured gap divided by ε varies by less than 6%, giving slopes 0.99 in ε and 0.49 in $\mathcal{E}_{\mathrm{temp}}$.

Temporal mass is not sufficient. Across four Markov descriptors and the operator sweep, transmitted energy predicts the collapse gap better at the input and first hidden layer: Pearson correlations are 0.694 and 0.921, against -0.023 and -0.606 for temporal mass. At layer two the two correlations are 0.923 and 0.951, so temporal mass is slightly better there. This is a qualified result, not a universal ordering. For realistic exponential filters with $\tau_{\mathrm{syn}} \in [0.25, 32]$, $\|\mathbf{B}\mathbf{\Pi}_\perp\|_{\mathrm{op}} > 1$ throughout and the exact DC condition never applies; the appropriate description is the finite-resolution family of Appendix A.2.12.



(a) collapse gap vs. α (b) middle band (c) Cholesky (d) coupled (e) the theorem's units

Figure 15: **Controlled collapse checks.** (a, b) Stability across membrane and reset settings; (c, d) local scaling exponents under the Cholesky and the coupled estimators; (e) the one-layer gap in the theorem's units.

A.4.4 Experiment III: exact temporal irreducibility

Exact pair. At $T = 2$, let P and Q assign probabilities $(0.1, 0.7, 0.1, 0.1)$ and $(0.1, 0.4, 0.4, 0.1)$ to $(00, 10, 01, 11)$. Their spike-count laws are identical, but a memoryless zero-threshold Gaussian layer gives correlations $\rho_P = 1/4$ and $\rho_Q = 1/5$. Hence $\Delta_{\text{orthant}} = (\arcsin(1/4) - \arcsin(1/5))/(2\pi)$ and $\Delta_{\text{rate,out}} = (\arcsin(1/4) - \arcsin(1/5))/(4\pi)$. The exact values, $8.168203203 \times 10^{-3}$ and $4.084101602 \times 10^{-3}$, are reproduced to relative error 2×10^{-15}. Monte Carlo with 2^{22} shared draws reaches deviations 1.6×10^{-5} and 2.8×10^{-6} (Figure 16).

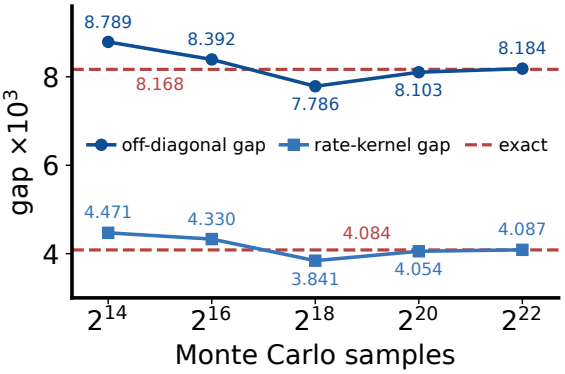


Figure 16: Monte Carlo convergence to the exact temporal and rate-kernel gaps of the $T = 2$ pair.

Equal-marginal pair. For the $T = 4$ construction of Theorem 3.5, the finite-width operator-norm gap falls from 0.267 at width 64 to 0.1123 at 65536, approaching the infinite-width value 0.11258. The rate contraction is too small to resolve at these widths: at width 16384 it is 2.4×10^{-3} with root-mean-square variation 1.2×10^{-2}. Exact cell probabilities instead give 4.406409×10^{-3}, within 8.6×10^{-8} of (8). We therefore use the operator norm, not the unresolved finite-width rate contraction, as the empirical separation statistic.

A.4.5 Experiment IV: count-preserving timing interventions

A temporal permutation $S_t^\pi = S_{\pi(t)}$ preserves each neuron's spike count exactly while changing its temporal covariance. We use global permutation, time reversal, adjacent-bin swaps and cross-neuron reshuffling. For each intervention we compare the raw difference $\Delta Q = \|\mathbf{Q} - \mathbf{Q}^\pi\|_F$, the scale-invariant transmitted difference between the correlation matrices of $\mathbf{BQB}^\top$ and $\mathbf{BQ}^\pi\mathbf{B}^\top$, and the output difference ΔK_{out}. Thresholds are calibrated separately in both arms to the same mean firing rate, so gain changes do not masquerade as timing effects.

Setup and result. The test comprises 63 cells at $T = 8$, crossing six encoders with identity, exponential, boxcar and random causal operators. Descriptors use 4096 input neurons and each output arm uses 2×10^5 shared Gaussian draws. All count-preservation assertions pass bitwise; both arms fire at mean rate 0.2500. Pooled over all cells, the scale-invariant transmitted difference explains ΔK_{out} at $R^2 = 0.827$ and Spearman 0.959, against $R^2 = 0.767$ for the raw difference and 0.272 for the unnormalized transmitted difference. On 36 held-out encoder–operator cells the corresponding R^2 values are 0.720, 0.628 and 0.109. Within an individual encoder the transmitted statistic is best in only three of six cases, so the evidence is pooled and qualified rather than universal (Figure 17). The i.i.d. Bernoulli exchangeability control has mean change 0.110, against 2.767 over the nonexchangeable cells.

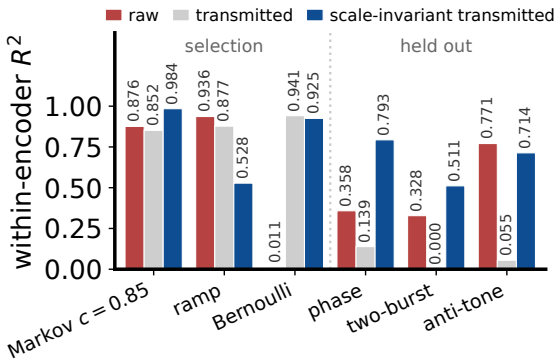


Figure 17: Within-encoder R^2 of the output change on raw, unnormalized transmitted and scale-invariant transmitted differences.

A.4.6 Experiment V: rate-null and rate-like tasks

We compare the full temporal kernel with the rate-projected kernel using the same kernel-ridge readout: $\widehat{\mathbf{Y}}_{\text{test}} = \mathbf{K}_{\text{test,train}}(\mathbf{K}_{\text{train,train}} + \lambda_{\text{KRR}}\mathbf{I})^{-1}\mathbf{Y}_{\text{train}}$. On handwritten digits we use $T = 8$, two hard-LIF layers, 200 training and 100 test examples, and compare exact time-to-first-spike, Bernoulli and persistent Markov encoders. Under the single-spike code every input neuron emits once, making the rate-projected kernel constant. Its accuracy is 0.110, near chance 0.100, whereas the full temporal kernel reaches 0.890. Bernoulli and Markov controls give no temporal advantage: full/rate accuracies are $0.900/0.930$ for Bernoulli and likewise agree for Markov. This experiment establishes the constructed rate null on one small dataset; it is not a broad benchmark claim.

A.4.7 Task alignment of temporal modes

Let $\mathbf{K}_{\text{temp}} = \mathbf{K} - \mathbf{K}_{\text{rate}}$ and $\mathbf{L} = \mathbf{Y}\mathbf{Y}^\top$. With $\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}^\top/N$, we measure $\mathcal{A}(\mathbf{K}, \mathbf{L}) = \langle \mathbf{HKH}, \mathbf{HLH} \rangle_F / (\|\mathbf{HKH}\|_F \|\mathbf{HLH}\|_F)$. Temporal mass asks whether timing variation exists; $\mathcal{A}(\mathbf{K}_{\text{temp}}, \mathbf{L})$ asks whether it aligns with the target. Appendix A.4.10 tests which quantity predicts the full kernel's advantage.

A.4.8 Experiment VI: conversion versus direct temporal training

Setup. Four networks share the same digit data, two width-128 hidden layers and mean-over-time readout: a ReLU network converted with per-layer activation percentiles, a hard-LIF network trained from scratch on Bernoulli rate codes, and the same hard-LIF network trained on time-to-first-spike or burst-onset codes. We evaluate three seeds, all 597 test images, both hidden layers, global time permutation and time reversal.

Result. Permutation sends the two timing-trained networks from 0.963 and 0.972 accuracy to chance and changes their logits by 0.92 and 0.99 of their norm. The converted and rate-trained networks change by at most 0.08, with accuracy changes below 0.01; time reversal gives the same separation. Transmitted energy does not classify the regimes: at layer two it is 0.63 for the converted network, above 0.30 and 0.23 for the timing-trained networks. The intervention supports the representation-use claim, whereas the proposed energy-only classifier fails (Figure 6).

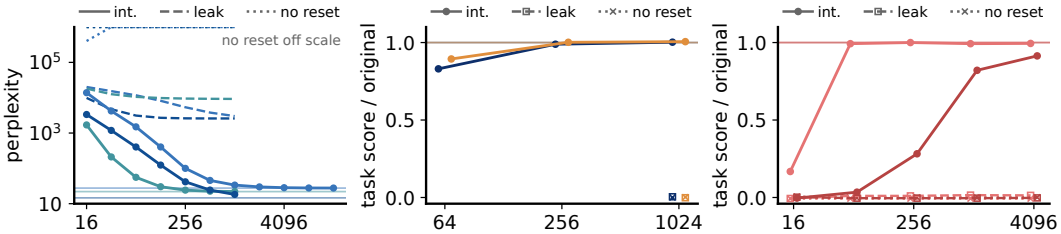
A.4.9 Converted checkpoints and exact rate-null checkpoints

Conversion protocol. Each activation entering a weight matrix becomes an integrate-and-fire population with zero initial state; signed activations use positive and negative populations. Per-channel thresholds are calibration maxima and are folded into the weights. Normalization, attention and gated linear units act on accumulated counts by emitting the increment of their cumulative analog output, so the cumulative output after t steps is exact. The rate endpoint is therefore the original checkpoint with activations clipped at their thresholds. The integrator $(\lambda, \rho) = (1, \vartheta)$ and the leak and no-reset controls run on the same evaluation set. Identity (46) and bound (47) are checked in double precision; test trajectories use single precision with TF32 disabled.

Compact vision, text and graph networks. Five two-layer width-256 ReLU networks cover MNIST, Fashion-MNIST, CIFAR-10 and two token-window tasks; three width-64 GCNs (Kipf & Welling, 2017) cover the Planetoid Cora, CiteSeer and PubMed splits (Yang et al., 2016). Their rate networks reproduce checkpoint outputs to numerical precision. At $T = 16384$, compact-network agreement is 0.9995–1.0000, and every GCN test prediction agrees. The identity residual is at most 2.0×10^{-8} for the compact networks and 7.3×10^{-9} for GCNs, with no bound violation. Leak errors are 0.38–0.57 for compact networks and 0.348–0.480 for GCNs; no-reset errors are 2.7–4.9 and 1.373–3.625.

Published vision checkpoints. Converted CIFAR ResNet-20/56 and CIFAR-100 ResNet-56 (He et al., 2016; Chen, 2019) retain accuracies 0.9260, 0.9439 and 0.7262 at $T = 16384$, with rate-network agreement 0.9989, 0.9994 and 0.9944. The ImageNet-1k ResNet-18 (Deng et al., 2009) uses thresholds from ten training images per class and is evaluated on all 50000 validation images. Its checkpoint and rate network both score 0.6976; the converted network scores 0.6979 at $T = 2048$, agrees with the rate network on 0.973 of images and has relative logit error 0.048. Across the checked vision pairs, identity residuals are at most 2.8×10^{-11} and no bound violation occurs. Both failed controls collapse to chance on the deep ResNets.

Language, speech, translation and control. OPT-125M/350M/1.3B (Zhang et al., 2022) rate networks reproduce checkpoint perplexities on WikiText-2 (Merity et al., 2017). At $T = 1024$, converted perplexities are 33.61, 22.11 and 18.26; extending OPT-125M to $T = 16384$ gives 27.723, and the first OPT-1.3B window is 0.22% above its rate network. Converted T5-small (Raffel et al., 2020) on WMT14 (Bojar et al., 2014) reaches BLEU (Papineni et al., 2002; Post, 2018) 24.09 at $T = 1024$, versus 24.04 for the checkpoint; S2T-small (Wang et al., 2020) on LibriSpeech (Panayotov et al., 2015) reaches word error rate 4.56%, versus 5.12%. Converted DQN agents (Mnih et al., 2015) for Pong and Breakout in the Arcade Learning Environment (Bellemare et al., 2013), evaluated with the Stable-Baselines3 protocol (Raffin et al., 2021), score 20.4 ± 0.6 and 344.8 ± 80.9 at $T = 4096$, versus 20.6 ± 0.5 and 377.1 ± 34.0. In every group the integrator error decreases with T, while leak and no-reset controls retain large error floors. Full trajectories are in Figure 18; main-text figures report the vision and graph trajectories.



(a) OPT, WikiText-2 (b) translation and speech (c) Atari control

Figure 18: **Sequence and control trajectories.** Translation, speech and Atari checkpoints under the integrator and failed controls.

Exact rate-null checkpoints. Six hard-LIF checkpoints are trained on single-spike, two-spike or order-reversal inputs whose rate quotient is exactly constant. Their same-weight rate surrogates return bitwise-identical logits on rate-equivalent inputs and under time permutation, with accuracies 0.0888, 0.1055, 0.1170, 0.0970, 0.0925 and 0.5000, exactly the predicted class frequencies. The checkpoints reach 0.9598, 0.9715, 0.9005, 0.8260, 0.3500 and 0.6515, and fall to 0.0787, 0.0938, 0.1125, 0.0955, 0.0950 and 0.4920 after time permutation. On Spiking Heidelberg Digits (Cramer et al., 2022), where counts do contain class information, the same-weight rate surrogate is exactly permutation invariant and scores 0.257; the full network scores 0.737 and falls to 0.271 after permutation.

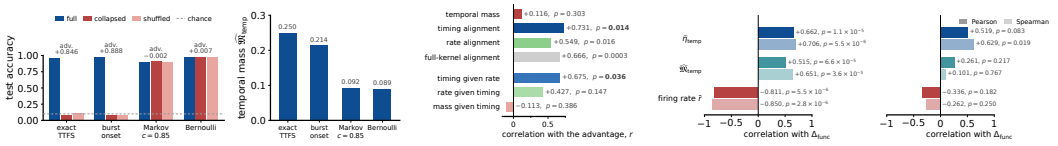
A.4.10 Trained finite SNNs

These experiments extend the diagnostics to trained finite networks; they do not verify the infinite-width theorem. We measure temporal mass, transmitted energy and the mean logit change $\Delta_{\text{func}} = N^{-1} \sum_{a=1}^{N} \|f(x^a) - f^{\pi}(x^a)\|_2$ under a count-preserving permutation. All diagnostics use forward spikes; surrogate derivatives are used only during training.

Encoder-family design: a negative identification result. Width-128, two-layer hard-LIF networks are trained on digits under rate and timing encoders and compared with a collapsed ReLU baseline and a shuffled-time SNN. The baselines are at chance for fixed-count timing codes and match the full network for rate codes. Across 12 encoders, the rate-versus-timing family indicator correlates with the full-network advantage at 0.997, and firing rates separate the two families without overlap. Although temporal mass correlates with advantage at 0.977, its exact permutation $p = 0.333$: this design cannot distinguish a diagnostic from the encoder family and supplies no diagnostic claim.

Mass versus task alignment. A second design separates the amount of temporal variation from its relevance. A fraction $\alpha \in \{0, 0.125, \ldots, 1\}$ of input neurons carries intensity in spike time, while $\beta \in \{0, 32, 64\}$ class-independent Markov units add temporal variation, yielding 27 trained networks. The timing-channel alignment predicts the full network's advantage at $r = 0.731$ with exact block-permutation $p = 0.014$, and remains significant after partialling out rate-channel alignment ($p = 0.036$). Raw temporal mass does not predict the advantage once alignment is known. Five independent repeats retain the timing-alignment result in every case ($p < 0.013$).

Transmitted energy versus functional change. On the same grid, first-layer transmitted energy correlates with Δ_{func}, but it also correlates with firing rate at -0.869. After conditioning on firing rate, its correlation with the logit change is -0.145 ($p = 0.537$), whereas firing rate conditioned on transmitted energy remains significant ($r = -0.635$, $p = 0.00026$). This is another negative result: in this diagnostic, where $\mu = 0$ and $\mathbf{B} = \mathbf{I}$, transmitted energy of binary spikes is determined by spike counts and does not isolate timing use beyond firing rate. This statement does not extend to a nonzero mean term or a general temporal operator.



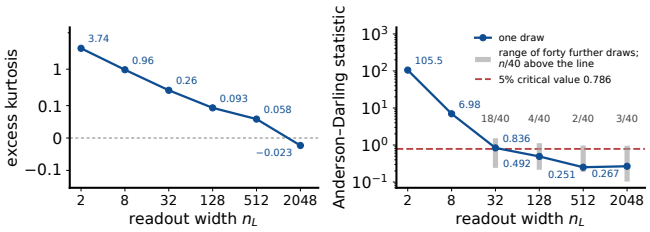
(a) accuracy, advantage (b) temporal mass (c) task alignment (d) hidden layer 1 (e) hidden layer 2

Figure 19: **Diagnostics on trained networks.** (a, b) Encoder-family confound; (c) task alignment on the α–β grid; (d, e) predictors of permutation-induced logit change.

A.4.11 Direct checks of auxiliary statements

Unless stated otherwise these checks use $T = 4$, one input, $(\lambda, \rho, \vartheta) = (0.5, 0.3, 0.35)$, $(\mu, \sigma) = (0, 1)$, zero initial state and identity temporal operator.

Gaussian readout. For 3000 independent readouts per width, excess kurtosis falls from 3.74 at width two to -0.023 at 2048. The Anderson–Darling statistic is below its five-percent critical value from width 128 onward in the displayed run; across 40 repeats, rejections at widths $128, 512, 2048$ occur $4, 2, 3$ times, close to the two expected under Gaussianity.



(a) excess kurtosis (b) Anderson–Darling test

Figure 20: Excess kurtosis and Anderson–Darling statistic of the finite-width readout.

Continuity, crossing and depth. Perturbing λ, ρ or ϑ gives fitted moment-map exponents 1.014, 0.977 and 0.984. The boundary-crossing probability falls from 0.151 to 0.0060 as ε decreases from 0.4 to 0.01; Lemma A.10's bound exceeds it by factors 23.9–51.8. The finite-depth recursion inequality holds at every layer of four-layer trajectories at $\varepsilon = 0.2$ and 0.05, with a wide but valid margin.

General witness and deficiency bound. Each of five temporal operators—identity, exponential, boxcar, random causal and random dense—maps a nonzero drive-correlation difference to a nonzero

output-moment difference; the respective maximum entrywise differences range over 0.085–0.300 and 0.020–0.048. For the exact rate-null target, the deficiency bound is attained for all five tested readouts: rate, complementary, two-bin, identity and single-bin. Their ridge risks equal the bounds $0.99929, 0, 0.99929, 0, 0.76667$, respectively.

A.4.12 Controls, protocol and predictions

Every intervention is checked to preserve per-example counts; arms are matched in firing rate or preactivation scale; and horizon, leak, reset, depth, width and temporal-operator alignment are varied separately. Random-network results use independent realizations, normalized Gram matrices for geometry comparisons and identical validation procedures for full and projected kernels. No task label enters temporal mass or transmitted-energy diagnostics.

Resolution and depth controls. Across the resolution sweeps, Δ^* correlates with the imposed filter or leak timescale at 0.895–1.000, but its log–log slope is below one and depends on tolerance; the data support ordering, not equality of scales. Direct perturbation of a fixed descriptor gives layer-map exponents 0.910–1.013, so the tested local map is Lipschitz. A full-versus-projected depth trajectory instead plateaus because it compares two different maps; it cannot estimate this modulus.

Predictions P1–P6. The decisive tests are: (P1) empirical temporal kernels converge with width; (P2) the full-versus-rate gap vanishes in the exact DC limit and decreases with transmitted temporal energy; (P3) equal-count inputs can yield different downstream kernels; (P4) architecture-transmitted differences predict output changes better than raw timing differences under matched scale; (P5) task advantage follows alignment rather than temporal mass alone; and (P6) rate-derived and timing-trained SNNs separate under count-preserving timing interventions. Table 3 reports each outcome, including the failed energy-only parts of P5–P6.

A.5 Numerical Support for Assumptions 1 and 2

Assumptions 1 and 2 are hypotheses of the theory, and this appendix reports what they cost in practice: whether they hold along descriptors the recursion actually produces, and with what constants. Everything here is computed in the model of (1), with a deterministic bias, so the rate-projected field is singular wherever the projector is, as the text intends.

Fixed-configuration genericity. For a fixed descriptor and operator, the following lemmas identify the thresholds at which boundary-nullity or local anti-concentration can fail. Throughout, $\boldsymbol{\ell}_t^\top$ is the t-th row of $\mathbf{L}_\lambda$, the box coordinates are $\mathbf{y} = \mathbf{L}_\lambda \mathbf{h}$ in each input block, and $\Theta_t = \{c_t(\mathbf{s}) : \mathbf{s} \in \{0,1\}^{t-1}\}$ is the set of reachable offsets of (14); the shift $c_t(\mathbf{s}) - \vartheta$ does not depend on ϑ.

Lemma A.31 (Generic boundary-nullity). *Fix $\lambda, \rho, \mathbf{B}$ and a descriptor c, and let $\mathbf{\Sigma}_y$ be the covariance of the box coordinates of the drive. Assumption 1 holds at every threshold ϑ outside the set*

$$\mathcal{N}(c) = \big\{\, \alpha_{a,t} - \big(c_t(\mathbf{s}) - \vartheta\big) \;:\; (a,t) \text{ with } (\mathbf{\Sigma}_y)_{(a,t),(a,t)} = 0,\ \mathbf{s} \in \{0,1\}^{t-1} \,\big\}, \tag{63}$$

where $\alpha_{a,t}$ is the mean of box coordinate (a,t), at which a coordinate of zero variance is almost surely pinned. In particular $|\mathcal{N}(c)| \le M(2^T - 1)$, so $\mathcal{N}(c)$ is finite and Lebesgue-null.

Proof. By Lemma A.2, the boundary of every cell lies in $\bigcup_{a,t} \bigcup_{c \in \Theta_t} \{\mathbf{z} : \boldsymbol{\ell}_t^\top \mathbf{z}_a = c\}$, a finite union of hyperplanes. A box coordinate with positive variance is a nonconstant one-dimensional Gaussian and charges no point, so its hyperplanes are null. A coordinate with zero variance equals $\alpha_{a,t}$ almost surely and charges a hyperplane if and only if $\alpha_{a,t} \in \Theta_t$, that is if and only if $\vartheta \in \mathcal{N}(c)$. Since $|\Theta_t| \le 2^{t-1}$, summing over t and over the M blocks gives the cardinality bound. □

Lemma A.32 (Anti-concentration with explicit constants). *With the notation above, write $\mathcal{D}$ for the pinned coordinates and $\mathcal{P}$ for the rest. If $\vartheta \notin \mathcal{N}(c)$, then Assumption 2 holds with the constant C_{ac} and any radius below the r_0 of*

$$C_{\mathrm{ac}} = \sum_{(a,t) \in \mathcal{P}} \frac{2\,|\Theta_t|\,\|\boldsymbol{\ell}_t\|_2}{\sqrt{2\pi}\,\sqrt{(\mathbf{\Sigma}_y)_{(a,t),(a,t)}}}, \qquad r_0 = \min_{(a,t) \in \mathcal{D}} \min_{c \in \Theta_t} \frac{|\alpha_{a,t} - c|}{\|\boldsymbol{\ell}_t\|_2}, \tag{64}$$

where $r_0 > 0$, with $r_0 = +\infty$ when $\mathcal{D} = \emptyset$, and any positive constant serves when $\mathcal{P} = \emptyset$.

Proof. The distance from $\mathbf{z}$ to $\{\boldsymbol{\ell}_t^\top \mathbf{z}_a = c\}$ is $|\boldsymbol{\ell}_t^\top \mathbf{z}_a - c|/\|\boldsymbol{\ell}_t\|_2$. For $(a,t) \in \mathcal{P}$ the one-dimensional Gaussian density bound gives $\mathbb{P}[|\boldsymbol{\ell}_t^\top \mathbf{Z}_a - c| \le r\|\boldsymbol{\ell}_t\|_2] \le 2r\|\boldsymbol{\ell}_t\|_2/(\sqrt{2\pi}\sqrt{(\boldsymbol{\Sigma}_y)_{(a,t),(a,t)}})$ for every $r > 0$. For $(a,t) \in \mathcal{D}$ the coordinate is almost surely $\alpha_{a,t}$, so $\mathbf{Z}$ lies at distance at least r_0 from each of its hyperplanes, and those hyperplanes contribute nothing for $r < r_0$. Summing the two families over the union bound gives (64). Positivity of r_0 is exactly $\vartheta \notin \mathcal{N}(c)$. □

Lemma A.33 (When a coordinate can be pinned)**.** *For $\sigma > 0$, box coordinate (a,t) is pinned if and only if $\mathbf{Q}_{aa}$ annihilates $\mathbf{A}^\top \boldsymbol{\ell}_t$. For the rate-projected field its variance is $\sigma^2(\boldsymbol{\ell}_t^\top \mathbf{A}\mathbf{v}_0)^2\, \mathbf{v}_0^\top \mathbf{Q}_{aa}\mathbf{v}_0$, so unless input a is silent ($\mathbf{v}_0^\top \mathbf{Q}_{aa}\mathbf{v}_0 = 0$) pinning requires $\boldsymbol{\ell}_t^\top \mathbf{A}\mathbf{v}_0 = 0$: a single linear equation on $\mathbf{A}$, hence a set of codimension one in the operator. At $t = 1$, and at every t when $\lambda = 0$, it says that the t-th row of $\mathbf{A}$ is orthogonal to $\mathbf{v}_0$. For $t \ge 2$ and $\lambda > 0$ it couples rows $1, \dots, t$, so a row $t \ge 2$ orthogonal to $\mathbf{v}_0$ pins its coordinate only under the further cancellation $\sum_{q<t} \lambda^{t-q}(\mathbf{A}\mathbf{v}_0)_q = 0$.*

For a fixed descriptor and operator, the three lemmas leave at most $M(2^T - 1)$ exceptional threshold values, and give the local constants in (64) elsewhere. This pointwise statement does not provide uniform anti-concentration constants or a common continuity modulus over a parameter or descriptor family; those remain explicit hypotheses of the finite-depth result.

Boundary counts and recursion sweep. The reachable-offset family in the three lemmas is smaller than the general bound in (38). Per input block, $\bar{J} = \sum_t |\Theta_t|$ equals T at $\rho = 0$, $2T - 1$ at $\lambda = 0$, $T(T+1)/2$ at $\lambda = 1$, and $2^T - 1$ for generic λ, rather than $T2^{T-1}$. Thus the generic count is smaller by a factor $T2^{T-1}/(2^T - 1) > T/2$, and the first three regimes grow only polynomially. At $\lambda = 0$, only the $q = t - 1$ reset term survives, so $c_t(\mathbf{s}) = \vartheta + \rho s_{t-1}$ has two values for $t \ge 2$ and one at $t = 1$; setting $\rho = 0$ recovers the memoryless regime used for the arcsine closed form. These counts are exact and were checked by enumeration, which gives the intermediate value 46 at $T = 6$, $\lambda = 1/\varphi$.

For the recursion diagnostic we sweep $(\lambda, \rho, \vartheta) \in \{0, 0.4, 0.8, 0.95, 1\} \times \{0, 0.5, 2\} \times \{-0.5, 0, 0.35, 1, 2\}$ at $T = 4$, with $M = 1$, $\mathbf{B} = \mathbf{I}$, $\mu = 0$, $\sigma = 1$, zero bias and zero initial membrane value. Each descriptor is taken two layers into the recursion initialized by independent Bernoulli(0.3) spikes, with each layer evaluated using 2^{18} quasi-random samples from one scrambling. Across the 75 configurations, both the full and rate-projected sampled maps have $0/75$ failures of Assumption 2. The two maps have $30/75$ and $14/75$ sampled zero-variance coordinates, inferred $r_0 \in [1, 2]$, and median estimated C_{ac} values 10.94 and 11.73, with no sampled coordinate pinned on a threshold. These are finite-sample diagnostics, not certificates of exact pinning or nondegeneracy in the limiting recursion.

The radius qualification is not removable. Extending the local bound to all radii replaces C_{ac} by $\max\{C_{\mathrm{ac}}, r_0^{-1}\}$. For a field at distance 1 with $\mathcal{E}_{\mathrm{temp}} = 10^{-6}$, $C_{\mathrm{ac}} = 1.6 \times 10^{-5}$ and $r_0 = \frac{1}{2}$, this enlarges the propagated constant by a factor 2500. At distance $\frac{1}{2}$, with $r_0 = \frac{1}{4}$ and $\mathcal{E}_{\mathrm{temp}} = 1.25 \times 10^{-7}$, the factor is 3969. In both cases the optimizing radius $(2\mathcal{E}_{\mathrm{temp}}/C_{\mathrm{ac}})^{1/3}$ equals r_0, so retaining the local qualification preserves the stated constant.

The three lemmas, measured. At $T = 4$, $\rho = 0.5$ and $\lambda = 0$, $|\Theta_t| = 1, 2, 2, 2$: seven offsets per input against the bound $2^T - 1 = 15$ of Lemma A.31. With $M = 1$, $\mu = 0$, $\mathbf{b} = \mathbf{0}$, $\sigma = 1$, the input population uniform on the silent and the all-spike path, so that $\mathbf{Q} = \frac{1}{2}\mathbf{1}\mathbf{1}^\top$, and $\mathbf{A} = \mathbf{I}$ with its second row made orthogonal to $\mathbf{v}_0$, the exceptional set computed from (63) is $\{-0.5, 0\}$, of cardinality two. At $\vartheta = 0.2$, (64) gives $C_{\mathrm{ac}} = 5.64$ from the three non-degenerate coordinates and $r_0 = 0.200$ from the pinned one, and $\mathbb{P}[\mathrm{dist} \le r]$, measured on one pseudo-random draw of 4×10^5 samples, is $0.0017, 0.0177, 0.0881$ at $r = 10^{-3}, 10^{-2}, 5 \times 10^{-2}$ against bounds $0.0056, 0.0564, 0.282$. Raising the leak removes this pinning: the number of zero diagonal entries of $\mathbf{L}_\lambda \mathbf{B}\boldsymbol{\Pi}\mathbf{Q}\boldsymbol{\Pi}\mathbf{B}^\top \mathbf{L}_\lambda^\top$ is one at $\lambda = 0$ and zero at $\lambda = 0.2, 0.5, 0.9, 1$, whereas a first row orthogonal to $\mathbf{v}_0$ stays pinned at every λ. Fifty thousand Gaussian draws of $\mathbf{A}$ produced no row orthogonal to $\mathbf{v}_0$ at $T = 2, 4, 8$.

What the two assumptions exclude, and how large that set is. Take $M = 1$, $T = 2$, $\lambda = \rho = u_0 = \vartheta = \mu = 0$, $\mathbf{b} = \mathbf{0}$, $\sigma = 1$, and the architecture $\mathbf{A}_\varepsilon = \mathbf{e}_1\mathbf{v}_0^\top + \varepsilon \mathbf{e}_2\mathbf{d}^\top$ with $\mathbf{v}_0 = 2^{-1/2}(1, 1)^\top$ and $\mathbf{d} = 2^{-1/2}(1, -1)^\top$, so $\|\mathbf{B}\boldsymbol{\Pi}_\perp\|_{\mathrm{op}} = \varepsilon$. Let the input population place mass $\frac{1}{2}$ on each of $(1, 1)$ and $(1, 0)$; the descriptor c is realizable. Because $\mathbf{d}^\top \boldsymbol{\Pi} = \mathbf{0}^\top$, the second row of $\mathbf{A}_\varepsilon \boldsymbol{\Pi}$ is *exactly* zero, so $\mathbf{Z}_{\mathrm{dc},2} \equiv 0$, which is exactly the threshold. Both assumptions fail here, and the moment gap $\|\mathcal{T}(c) - \mathcal{T}_{\mathrm{dc}}(c)\|_\infty$ of Theorem A.11 is $1/2$ for every ε.

The two assumptions are not independent: the configuration requires two exact algebraic coincidences. The second coordinate of the drive satisfies $Z_2 \sim \mathcal{N}(0, \varepsilon^2/4)$ while $\mathbf{Z}_{\mathrm{dc},2} \equiv 0$, so the two arms disagree on it with probability $\mathbb{P}[\Psi(\mathbf{Z})_2 \neq \Psi(\mathbf{Z}_{\mathrm{dc}})_2] = \Phi(-\frac{2|\vartheta|}{\varepsilon})$, which tends to zero for every fixed $\vartheta \neq 0$ and equals $\frac{1}{2}$ only at $\vartheta = 0$. Measured with one pseudo-random draw of 4×10^6 samples under the coupling (34)–(35), the gap at $\vartheta = 10^{-3}, 10^{-2}, 10^{-1}$ falls to 0.0227, 3.2×10^{-5} and 0 by $\varepsilon = 10^{-3}$, 5×10^{-3} and 2×10^{-2} respectively, and to zero at every one of them by $\varepsilon = 2 \times 10^{-4}$; only the row $\vartheta = 0$ stays at $1/2$. The architectural coincidence is equally thin: a row of $\mathbf{A\Pi}$ vanishes precisely when that row of $\mathbf{A}$ is orthogonal to $\mathbf{v}_0$, a codimension-one condition, and $20{,}000$ Gaussian draws of $\mathbf{A}$ produced none.

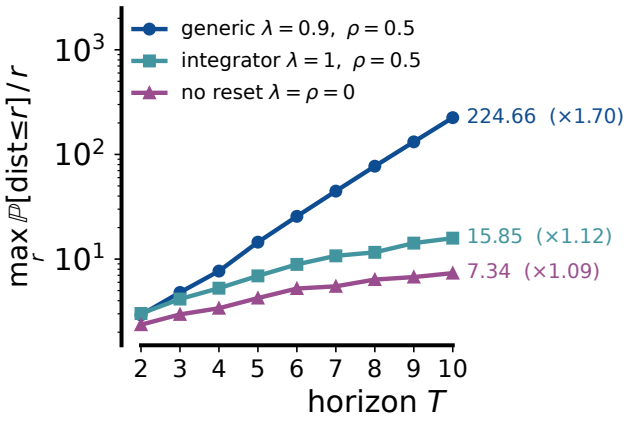


Figure 21: Measured anti-concentration constant against the horizon T in three regimes.

Scaling in the horizon. Measuring $\max_r \mathbb{P}[\mathrm{dist} \le r]/r$ directly at $\vartheta = 0.3$, over $r \in \{10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$ with 4×10^5 samples at each $T = 2, \ldots, 10$, for $M = 1$, zero initial membrane value and a centred field whose box coordinates have covariance $\mathbf{L}_\lambda \mathbf{G}\mathbf{G}^\top \mathbf{L}_\lambda^\top / T + \frac{1}{4}\mathbf{I}_T$, with $\mathbf{G}$ a standard Gaussian $T \times T$ matrix, $\mathbf{G}$ and then the samples being drawn from one stream of standard normals restarted for every T and every regime, and dist the distance of $\mathbf{y}$ to the union of all hyperplanes $\{y_t = c\}$, $c \in \Theta_t$, gives the three curves of Figure 21, matching the three regimes above; each curve is labelled at $T = 10$ with its value there and, in parentheses, the ratio of the values at $T = 10$ and $T = 9$. The generic constant grows geometrically, so a bound that unions over the whole boundary set is non-vacuous only over a window that closes as the horizon grows; the local linear behavior observed in Section A.2.7 does not remove this boundary-count dependence from the proved general bound.