WHAT DOES A FITTED MODEL REPRESENT? GRADED ORDERS AND SAMPLING COST IN ECHO-STATE MODEL SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Representing a time series by the parameters of a model fitted to it is a common approach, from model-space learning to weight-space learning, yet little is known about what such a representation retains. We analyse echo-state networks, in which a ridge readout of a fixed random reservoir is fitted to forecast each sequence and serves as its representation. The fitted readout is a statistic of the law of the process. Its linear part depends only on the autocovariance, and higher-order structure enters order by order: for scalar input, once lower-order moments agree, a difference of order $2k+2$ enters at order λ^{2k+1} in the input scale λ, generically when it is a joint cumulant of the forecasting target and the current input. Detecting it from one trajectory of length L is governed by $\lambda^{4k}L$, which we establish for bounded moving averages at a fixed ridge with a shared non-degenerate linear noise as a local Gaussian limit, with a lower bound on the error of every classifier of bounded complexity that population LDA attains. The same analysis reveals blind spots created by the forecasting target and by the symmetry of the reservoir, and shows how to remove them. Together, these results explain when a fitted-model representation distinguishes time series from different processes, and what to change when it does not.

1 INTRODUCTION

Representing a data point by the parameters of a model fitted to it is a recurring idea in machine learning. Time series have long been represented by fitted autoregressive coefficients (Piccolo, 1990; Maharaj, 1996), and model-space methods fit a reservoir or another dynamical model to each sequence and classify the fitted parameters (Chen et al., 2014; 2015; Aswolinskiy et al., 2018; Bianchi et al., 2021). Recent work extends model spaces to spectral dynamics (Liu et al., 2025), to multivariate series through graphs (Liu et al., 2026), to irregularly sampled sequences (Chen et al., 2025; Zhou et al., 2026), and to fault and subsurface diagnosis (Chen et al., 2024; Zhou et al., 2025). Fisher scores use the gradient of a fitted likelihood (Jaakkola & Haussler, 1999). Weight-space learning represents images, shapes and models by the parameters of networks fitted to each of them (Dupont et al., 2022; Schürholt et al., 2022), and dedicated architectures learn from such weights (Navon et al., 2023; Lim et al., 2024; Zhou et al., 2024). What such a representation retains, and at what cost, has little theory (Qiu et al., 2026).

A fitted model summarizes how a sequence evolves, so sequences from one process lead to nearby models even when the sequences themselves are far apart (Figure 1). Whether the models of different processes separate depends on what the fit retains. Two processes can share their spectrum and their marginal law and still differ, as a process and its time reversal do in a fourth-order lagged moment (§2), and every representation that depends on the data only through the autocovariance maps both to one value. A fitted nonlinear model can tell such processes apart, but it is not clear which differences it retains, and how long a trajectory it needs before a retained difference rises above the sampling noise of the fit.

We answer these questions for echo-state networks (Jaeger, 2001). Every sequence drives the same fixed random recurrent network, and a ridge readout fitted to forecast the sequence is its representation. The fit is linear and in closed form while the features are nonlinear, so the representation captures

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

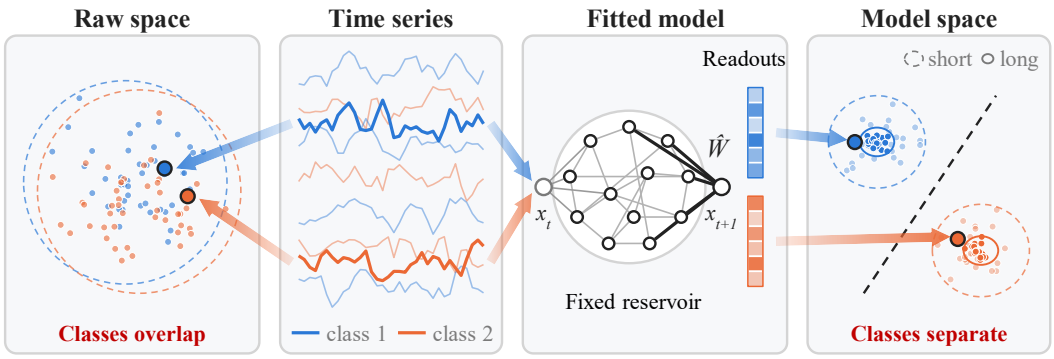


Figure 1: Learning in the model space (schematic). *Raw space:* in the raw signal space, time series from different processes can be hard to separate. *Time series:* sequences of the two classes, each one point in the raw space. *Fitted model:* each sequence drives the same fixed reservoir, and a linear readout $\hat{W}$ is fitted to predict the sequence. *Model space:* each sequence is then represented by its fitted readout, one point in this space. Readouts of one class describe the same dynamics, so they gather together and apart from the other class, more tightly as the sequences grow longer (dashed: short, solid: long). The highlighted pair is close in the raw space and far apart in the model space. Proposition B.1 makes this precise for moving averages.

more than second-order statistics and remains tractable, as random-feature models do (Rahimi & Recht, 2007). The fitted readout converges to a functional of the law of the process, so it is a statistic of the law rather than of the individual sequence. Its linear part depends only on the autocovariance. For scalar input, a difference in a fourth-order joint cumulant of the target and the current input enters at order λ^3 in the input scale λ and, for bounded moving averages at a fixed ridge with a shared non-degenerate linear noise, needs a trajectory of length of order λ^{-4} to be detected, and each further order costs two more powers of λ in the readout and four in the length. The target and the symmetry of the reservoir decide which differences can enter. The restriction to the conditional mean uses only that the features are functions of the past, and the sign symmetry only that they are odd, so both hold for other fitted models, whereas the graded orders and the sampling law use the echo-state map.

Earlier expansions of reservoir states describe how a nonlinearity mixes cumulants into the state (Nestler et al., 2020). We study instead the readout fitted to each trajectory, as a statistic with its own sampling law (§6).

Contributions.

- *Graded visibility* (§3). Once the lower-order joint moments of two scalar processes agree, a difference of order $2k+2$ enters the readout at λ^{2k+1}, and it does so for generic reservoirs when the joint cumulant $c_k(s)$ of the target with $2k+1$ copies of the current input differs (Theorem 3). On finite-memory reservoirs, where the separation is computed exactly, the differences become detectable at order-one scales set by their order and the size of this cumulant.
- *Sampling cost* (§4). For bounded moving averages at a fixed ridge and under a shared non-degenerate linear noise, such a difference is detectable from one trajectory on the scale $\lambda^{4k}L$, so that halving the scale costs sixteen times the length at fourth order. We prove a lower bound on the error of every classifier of bounded complexity that population LDA attains (Theorem 4).
- *Blind spots and remedies* (§2, §5). At every scale and ridge, the population next-step readout depends on the target only through its conditional mean, and an odd reservoir gives a process and its reflection one law of the readout. A bias together with a squared target restores the missing signal.

2 The representation and the role of the target

We are given N labeled sequences with values in $\mathbb{R}^{d_x}$. Stage one fits one ridge readout per sequence on a shared reservoir, and stage two trains a classifier on the fitted readouts, so labels enter only at stage two. The reservoir is an echo-state network of width n (Jaeger, 2001; Yildiz et al., 2012;

Lukoševičius, 2012), with a fixed random matrix W of spectral radius $\rho(W) < 1$, a fixed input matrix W_{in} and an input scale $\lambda > 0$. Its state follows

$$h_t = \tanh\big(W h_{t-1} + \lambda W_{\text{in}} x_t\big), \qquad h_0 = 0. \tag{1}$$

The readout is fitted to the target $g_s(x)_t = x_{t+s}$, the sequence s steps ahead, and we call s the span. The next-step target $s = 1$ is the one used in practice (Chen et al., 2014). The self-step target $s = 0$ is convenient for analysis. After a washout of t_{w} steps the fit window is $I_s = \{t_{\text{w}} + 1, \dots, T - s\}$, of length L, and the sample moments and the ridge readout are

$$\hat{\Sigma}_h = \tfrac{1}{L} \sum_{t \in I_s} h_t h_t^\top, \qquad \hat{M} = \tfrac{1}{L} \sum_{t \in I_s} x_{t+s} h_t^\top, \qquad \hat{W} = \hat{M} \, (\hat{\Sigma}_h + \gamma I_n)^{-1}, \tag{2}$$

with ridge penalty $\gamma > 0$. The representation is the vector of readout weights, usually L2-normalized before classification. The ridge is set either at a fixed level or relative to the state energy of the sequence, $\gamma = g \operatorname{tr}(\hat{\Sigma}_h)/n$, and we use both. Table 3 collects the notation.

Assumption 1 (Standing setup)**.** *The reservoir (W, W_{in}) is fixed, with $\rho(W) < 1$, and the ridge penalty is $\gamma > 0$. Each class y is a zero-mean, strictly stationary and ergodic process P_y with finite second moments, and every sequence of class y is a trajectory of P_y. Results that need more, such as a contraction or bounded innovations, state it.*

Assumption 2 (Contraction)**.** *There exist a diagonal $D \succ 0$ and $q < 1$ with $W^\top D W \preceq q^2 D$. The special case $\|W\|_2 \le q < 1$, with $D = I$, suffices, and the orthogonal family $W = \rho Q$ satisfies it with $q = \rho$ (Remark E.1 treats other families).*

As $L \to \infty$ the fitted readout of a class-y sequence converges to the population readout

$$\bar{W}_{g,y} := M_{g,y} (\Sigma_{h,y} + \gamma I)^{-1}, \qquad \Sigma_{h,y} = \mathbb{E}_y[h_t h_t^\top], \qquad M_{g,y} = \mathbb{E}_y[x_{t+s} h_t^\top], \tag{3}$$

where the moments are taken at the stationary state, which the trajectory started at $h_0 = 0$ reaches under the contraction of Assumption 2 (Proposition A.1). We write $\Delta A := A_1 - A_2$ for the class difference of any class-indexed quantity.

This limit distinguishes the representation from the raw sequence. The population readout is a function of the law of the process, and the fitted readout of one sequence converges to it, with fluctuations of order $1/\sqrt{L}$ for the moving averages of §4 (Lemma D.1). A raw sequence has no such limit. The distance between two raw sequences is dominated by their innovations, so two sequences of one class are about as far apart as two sequences of different classes. For moving averages of symmetric innovations that share their autocovariance, the error of Euclidean k-nearest-neighbour rules on the raw sequences tends to $\frac{1}{2}$, while nearest neighbours on the fitted readouts become consistent from one training sequence per class whenever the population readouts differ (Proposition B.1).

Linearizing the recurrence gives a readout that depends on the process only through its autocovariance (Proposition C.1). What the nonlinearity adds depends first on the target.

Corollary 1 (The next-step readout sees the conditional mean)**.** *Let Assumption 2 hold and $s \ge 1$. The stationary state h_t is a function of the past inputs $x_{\le t}$, and*

$$M = \mathbb{E}\big[\, \mathbb{E}[x_{t+s} \mid x_{\le t}] \; h_t^\top \big]. \tag{4}$$

The target therefore enters the population readout only through its conditional mean $\mathbb{E}[x_{t+s} \mid x_{\le t}]$, jointly with the law of the state. If the conditional mean vanishes, as for martingale differences, the population readout is zero at every scale, every ridge and for every such reservoir.

Proof. Under Assumption 2 the stationary state is the unique causal solution of (1), a measurable function of $x_{\le t}$ (Proposition C.4). The tower property gives (4). If the conditional mean vanishes, then $M = 0$ and $\bar{W} = M(\Sigma_h + \gamma I)^{-1} = 0$. □

At the next-step target, two iid classes, and more generally any two classes of martingale differences, therefore have population readout zero, whatever their marginal laws. This concerns the centre of the representation. The fitted readouts can still differ in law, for instance in covariance. The sign symmetry below is stronger and concerns the whole law.

Remark 2 (Sign symmetry). *With $h_0 = 0$, no bias inside the* tanh *and no readout intercept, $h_t(-x) = -h_t(x)$ for every t, so a process and its reflection $x \mapsto -x$ induce one law of the fitted readout, and no classifier of the readout separates them (Remark E.2).*

Each blind spot has a simple remedy. A bias inside the tanh breaks the sign symmetry. A difference in volatility is invisible to the next-step readout, since only the conditional mean of the target enters it. Adding the squared target x_{t+1}^2 as a second target makes the conditional variance a quantity being predicted. Without a bias, however, the state is an odd function of the input path and x_{t+1}^2 is even, so their product averages to zero under any law symmetric under $x \mapsto -x$. A difference in conditional variance between such classes therefore needs both remedies, which §5.2 tests.

We use one family of examples throughout, a process and its time reversal. Let x_t be a scalar stable linear filter of iid symmetric innovations, and let the second class use the time-reversed filter. Reversal preserves the autocovariance and the marginal law (Weiss, 1975; Ramsey & Rothman, 1996). The two-tap pair has coefficients $(1, \frac{1}{2})$ and innovations ± 1, and the eight-tap pair has coefficients $0.85^j \cos(2\pi \cdot 0.11 j)$ for $j = 0, \ldots, 7$. The lag-two pair compares $x_t \propto e_t + \frac{1}{2} e_{t-2}$ with its reversal, whose difference lies two steps back (Appendix F).

3 Statistical order and visibility

The mechanism is visible on one unit without memory. For $h_t = \tanh(\lambda w x_t)$, the expansion $\tanh u = u - \frac{1}{3}u^3 + \cdots$ gives the cross-moment with the target

$$M = \lambda w \, \mathbb{E}[x_{t+s} x_t] - \tfrac{1}{3}\lambda^3 w^3 \, \mathbb{E}[x_{t+s} x_t^3] + \cdots , \tag{5}$$

so the term of order λ^{2k+1} carries moments of order $2k+2$. The first possible class difference beyond the autocovariance is therefore cubic, through a fourth-order cross-moment with the target, and whether it appears depends on that cross-moment. On a reversal pair at the next-step target, $\mathbb{E}[x_{t+1} x_t]$ is shared and $\mathbb{E}[x_{t+1} x_t^3]$ differs by the joint cumulant $\Delta\mathrm{cum}(x_{t+1}, x_t, x_t, x_t)$, which is -0.48 on the two-tap pair (Appendix F). The following theorem makes this exact for a reservoir with memory and says when the difference appears.

Theorem 3 (Graded orders at a fixed span). *Let Assumption 2 hold, let $d_x = 1$, fix a span $s \geq 0$, and let P, P_1 and P_2 be zero-mean and strictly stationary with finite moments of all orders. Write $\tau_{2k+1} \neq 0$ for the Taylor coefficient of order $2k+1$ of* tanh.

(a) Orders. *For every $K \geq 0$, the cross-moment $M_{g_s,P}(\lambda)$ of* (3) *under P at scale λ satisfies*

$$M_{g_s,P}(\lambda) = \sum_{j=0}^{K} \lambda^{2j+1} M_{g_s,P}^{(2j+1)} + O(\lambda^{2K+3}), \tag{6}$$

where $M_{g_s,P}^{(2j+1)}$ is a linear function of the joint moments of P of order $2j+2$. If P_1 and P_2 share every joint moment of order at most $2k$, then $\Delta\Sigma_h = O(\lambda^{2k+2})$ and

$$\Delta\bar{W} = \gamma^{-1} \lambda^{2k+1} \Delta M_{g_s}^{(2k+1)} + O(\lambda^{2k+3}). \tag{7}$$

(b) Visibility. *The coefficient $\Delta M_{g_s}^{(2k+1)}$ is defined for every W with $\rho(W) < 1$. Let*

$$c_k(s) := \Delta \, \mathrm{cum}\big(x_{t+s}, \underbrace{x_t, \ldots, x_t}_{2k+1 \text{ copies}}\big). \tag{8}$$

At $W = 0$ the coefficient is $(\Delta M_{g_s}^{(2k+1)})_i = \tau_{2k+1} \, c_k(s) \, w_i^{2k+1}$, with w_i the i-th entry of W_{in}. If $c_k(s) \neq 0$, then $\Delta M_{g_s}^{(2k+1)} \neq 0$ for Lebesgue-almost every (W, W_{in}) with $\rho(W) < 1$, and along $W = \rho Q$ with Q orthogonal and $W_{\mathrm{in}} \neq 0$ fixed, for all but countably many $\rho \in (0, 1)$.

Proof in Appendix C.2.1.

For $k = 1$ the theorem says that a difference in fourth-order moments enters at λ^3, as in (5). The factor γ^{-1} appears because at small scale $\Sigma_h = O(\lambda^2)$ is negligible against the ridge, so that the

readout is the cross-moment divided by γ. Part (a) says that a difference in moments of order $2k+2$ cannot enter before λ^{2k+1}, and part (b) says when it does. The witness $c_k(s)$ is a cumulant with the target, so the target affects which differences are visible. The condition $c_k(s) \neq 0$ is sufficient rather than necessary. A reservoir with memory also weights lagged cumulants, and it separates the lag-two pair, for which $c_1(1) = 0$ (see below). Two iid classes that differ in their fourth moment have $c_1(1) = 0$, and their population next-step readouts are both zero, as Corollary 1 requires.

Beyond the order, the memory of the reservoir decides which lags enter. At cubic order the class difference is a fixed linear image of the fourth-order lag moments $\mathbb{E}[x_{t+s}x_{t-j_1}x_{t-j_2}x_{t-j_3}]$, and a reservoir with $\|W\|_2 \leq q$ weights a lag triple that reaches j steps back by $O(q^j)$ (Proposition E.3). With a finite memory the description is exact at every scale. If the graph of W has no path longer than p and the input takes finitely many values, the fitted readout is at every scale a ridge regression of x_{t+s} on indicators of the last $p+1$ inputs, with a penalty that depends on the scale (Proposition D.6). A reservoir without memory therefore has one population readout for both classes of the lag-two pair, since its state depends on the current input alone and consecutive values of this pair have one joint law in both classes.

4 Sampling cost

Visibility alone does not decide whether one trajectory suffices. Suppose the class difference enters the population readout at order λ^{2k+1}. The leading sampling fluctuation comes from the linear term and has scale $\lambda/\sqrt{L}$, so the squared ratio is proportional to $\lambda^{4k}L$. To make this exact, let $h_t^{(1)}$ be the linearized state, $\hat{M}_1$ its sample cross-moment with the target and V_1 the long-run covariance of $\hat{M}_1$,

$$h_t^{(1)} = \sum_{j\geq 0} W^j W_{\text{in}} x_{t-j}, \qquad \hat{M}_1 = \frac{1}{L}\sum_{t\in I_s} x_{t+s} h_t^{(1)\top}, \qquad V_1 := \lim_{L\to\infty} L\,\mathrm{Cov}(\hat{M}_1). \tag{9}$$

For integers $H, \nu \geq 1$, let $\mathcal{A}_{H,\nu}$ be the family of sets defined by a Boolean function of at most H conditions $p > 0$, with polynomials p of degree at most ν.

Theorem 4 (Sampling law and lower bound)**.** *Let Assumption 2 hold, let $d_x = 1$, fix $\gamma > 0$ and a span s, and let each class be a moving average $x_t = \sum_{j=0}^{m} \vartheta_j e_{t-j}$ of iid, centred and bounded innovations. Let the two classes share every joint moment of order at most $2k$ for some $k \geq 1$, let $\delta_k := \mathrm{vec}\,\Delta M_{g_s}^{(2k+1)} \neq 0$, and let them share V_1, with $V_1 \succ 0$. Sharing is automatic for $k \geq 2$, since V_1 depends only on joint moments of order at most four. Let $L \to \infty$ and $\lambda \to 0$ with $\lambda^{4k}L \to \zeta \in [0,\infty)$.*

(a) Local limit. *With $\langle \hat{W} \rangle_L := \frac{1}{2}(\mathbb{E}_1\hat{W} + \mathbb{E}_2\hat{W})$ the mean of $\hat{W}$ over the two classes,*

$$\frac{\sqrt{L}}{\lambda}\big(\hat{W} - \langle \hat{W} \rangle_L\big) \;\Rightarrow\; \mathcal{N}\big(\pm\tfrac{1}{2}\sqrt{\zeta}\,\gamma^{-1}\delta_k,\; \gamma^{-2}V_1\big) \tag{10}$$

in law, with the sign $+$ on the first class and $-$ on the second.

(b) Lower bound. *Consider any classifier of $\hat{W}$ that is trained on data independent of the test sequence and whose decision region for one class is convex or lies in $\mathcal{A}_{H,\nu}$, for fixed H and ν. Under equal class probabilities its test error is at least*

$$\Phi\big(-\tfrac{1}{2}\sqrt{\zeta\kappa_k}\big) - o(1), \qquad \kappa_k := \delta_k^\top V_1^{-1}\delta_k, \tag{11}$$

with Φ the standard normal distribution function, and LDA with the population parameters attains this value in the limit.

Proof in Appendix D.6.

Part (a) is a central limit theorem under local alternatives (van der Vaart, 1998), in which the scale sets the distance between the two classes. Part (b) places the loss in the representation rather than in a particular classifier. The constant κ_k is the squared length of the signal δ_k in the metric of the linear noise, and once $\zeta\kappa_k$ is small no rule of bounded complexity is much better than chance, whatever independent data it is trained on.

For a fourth-order difference, $k = 1$, halving the input scale requires sixteen times the length to keep the limiting separation. Figure 2 shows this on the two-tap pair, with $\kappa_\star := \kappa_1$. Below a crossover

scale, the measured separation lies on $\lambda^4 L\kappa_\star$ and LDA trained on 1000 sequences per class follows the limit of part (b). Above it the separation grows more slowly than the law.

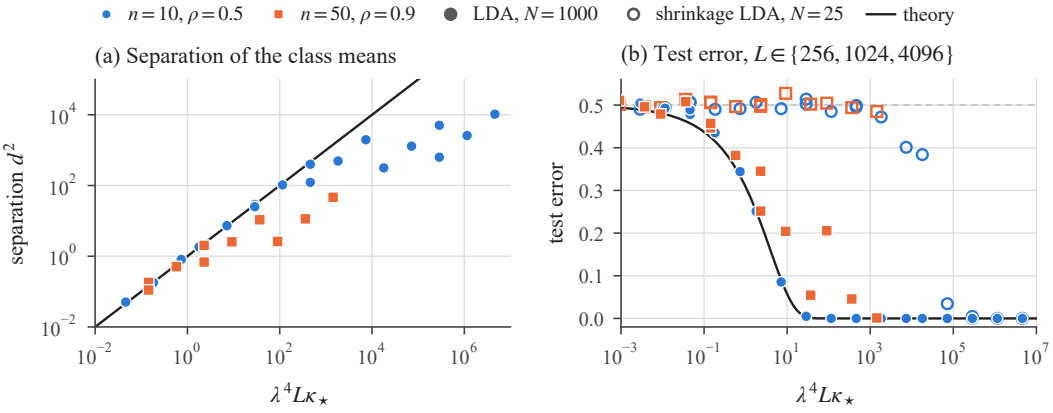


Figure 2: The sampling law on the two-tap pair (§4), tanh readout at the fixed ridge $\gamma = 1$, $L \in \{256, 1024, 4096\}$ and λ from 0.01 to 1, on two reservoirs: circles $n = 10$, $\rho = 0.5$, squares $n = 50$, $\rho = 0.9$. *(a)* Measured separation of the class means against $\lambda^4 L\kappa_\star$, with the diagonal of the theorem (solid line). The points leave the diagonal where the higher-order terms set in. *(b)* Test error of LDA trained on 1000 sequences per class (filled) and of shrinkage LDA trained on 25 (open), with the limit $\Phi(-\frac{1}{2}\sqrt{\lambda^4 L\kappa_\star})$ (solid line). The filled squares above the limit lie beyond the crossover of panel (a). The open markers stay at chance where the population rule separates the classes, which is the cost of estimating the metric from few sequences (§4, Appendix F.6).

The bound covers linear rules, also after an affine map and L2 normalization, QDA, trees, forests, ReLU networks of fixed size and nearest-neighbour rules with a fixed number of training sequences, but not rules whose complexity grows with L (Appendix D.5).

The constant $\kappa_\star$ is large when the signal has a component in directions where the linear noise is small, outside the leading Krylov directions $W^j W_{\text{in}}$ (Proposition D.5). A classifier trained on few sequences estimates these directions poorly, and shrinkage toward the identity discards them. In Figure 2b shrinkage LDA with 25 sequences per class stays at chance over more than an order of magnitude of $\lambda^4 L\kappa_\star$ in which LDA on 1000 sequences has error near zero. The length L sets the noise of each readout, and the number N of training sequences sets how well the class means and the metric are estimated. Theorem 4 concerns the first, and its bound is attained by LDA with the population parameters, which does not guarantee that a rule trained on N sequences attains it.

The theorems describe the small-scale end of the scale axis. With a unit-norm W_{in} and unit-variance input, λ is the size of the input drive relative to the range in which tanh is close to linear, so they hold for $\lambda \ll 1$. At the other end, as $\lambda \to \infty$, the state tends to $\text{sign}(W_{\text{in}} x_t)$ with or without memory, and for scalar input the readout reduces to one sign statistic with its own sampling law (Appendix D.4).

Held-out prediction selects order-one and larger scales (Appendix F.7). There several orders contribute at once, and the separation is no longer a single power of λ. On reservoirs of finite memory, where it can be computed at every scale, the onset of detection lies at order-one scales that depend on both the order of the difference and the size of its witness, and the leading term places it more accurately the longer the sequence (§5.1).

5 Experiments

Unless stated otherwise, reservoirs are $W = \rho Q$ with Q Haar orthogonal. Every data-dependent choice is made on the training sequences alone (Appendix F).

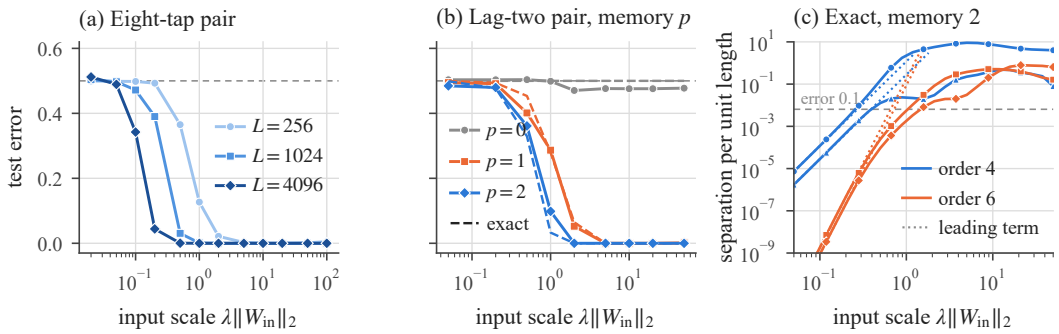


Figure 3: Scale and memory (§5.1). *(a)* Shrinkage-LDA error ($N = 25$) of the fitted $\tanh$ readout against the scale, eight-tap pair, $\rho = 0.9$, relative ridge chosen by held-out R^2. Dashed grey: chance. *(b)* Lag-two pair, memory p, $L = 256$: simulated error (markers, mean over three draws) and population LDA error from the exact moments (dashed). At $p = 0$ the simulated error dips below $\frac{1}{2}$ because the class covariances differ (Appendix F.8). *(c)* Exact separation per unit length of shrinkage LDA with the population parameters, on reservoirs of memory two ($n = 24$, fixed ridge $\gamma = 1$, direction from the pooled covariance shrunk toward its trace with weight 0.1), two-tap pairs with $\vartheta_1 = \frac{1}{2}$ whose first difference is at order four (blue) or six (orange), circles and triangles for ± 1 and three-point innovations, squares and diamonds for Gauss–Hermite and four-point (Appendix F.10), geometric mean over three reservoirs. Dotted: the leading term, fitted below $\lambda = 0.12$. Dashed grey: Gaussian error 0.1 at $L = 1024$ in a shared covariance.

5.1 Orders and sampling cost

The measured readout has the orders of Section 3. With the lower-order terms subtracted as control variates computed exactly, the log-log slopes of the cubic term at both targets and of the quintic term at the self-step target are within 0.017 of 3 and 5 (Appendix F.4). At sixth order, an exact computation on one memoryless unit recovers the constant of the sampling law, $\kappa_2 = \frac{576}{23125}$, to four digits (Appendix D.6).

Panel (a) of Figure 3 follows the readout across the scale on the eight-tap pair with 25 training sequences per class, at the relative ridge used in practice, for which Remark D.4 gives the leading terms and the conditions of the $\lambda^4 L$ law. It is at chance at small scale, as Figure 2b explains, and separates from λ of order one. Panel (b) shows the lag-two pair on reservoirs of finite depth. The error of the population rule computed exactly from Proposition D.6 is within 0.07 of the simulated error, and the reservoir without memory stays at chance (Appendices F.7 and F.8).

Panel (c) locates the onset of detection. On reservoirs of memory one and two, Proposition D.6(c) gives the population readouts and their long-run covariances at every scale, here for two-tap pairs that differ first at fourth or at sixth order, and hence the exact Gaussian error of shrinkage LDA with the population parameters. At small scale the log-log slopes of its separation are 4 and 8 to within 0.07, from the orders of Theorem 3. This Gaussian error at $L = 1024$ falls to 0.1 between $\lambda = 0.17$ and 4.4, depending on the order and on the size of the witness. The leading term, extrapolated to this error, gives the onset at fourth order within a factor 0.78 to 3.3, and at sixth order it comes too early, by a factor 1.5 to 4.3. Over the 48 cases the factor is 0.71 to 5.9 at $L = 256$, 0.78 to 4.3 at $L = 1024$ and 0.83 to 2.8 at $L = 4096$, narrowing as the sequences grow (Appendix F.10).

Finally, the order of a difference also decides which fixed statistics detect it. When the classes differ first at sixth order, twenty fourth-order reversibility statistics stay at chance, while the unchanged readout, whose separation grows as $\lambda^8 L$ at small scale since $c_2(1) \neq 0$, has error 0.48, 0.25, 0.02 and 0.00 at $L = 4096$ and $\lambda = 0.5$, 1, 2 and 4 (Appendix F.9). Statistics built for the right order and lags are more efficient per unit length, by 1.4 to 20 times on the two-tap and eight-tap pairs (Appendix F.11), but each covers only the order and the lags it was built for.

Table 1: The controlled pair of martingale differences, $L = 1024$, 50 training sequences per class, scale chosen without labels, four reservoir draws. Error: shrinkage-LDA test error, mean over draws. Mean gap: largest class-mean difference of a readout coordinate at $\lambda = 1$, in pooled standard deviations, range over draws. The squared target is fitted alongside the original one.

n		plain	bias	x^2 target	bias and x^2 target
10	error	0.502	0.500	0.493	0.000
	mean gap	0.07–0.12	0.09–0.15	0.07–0.15	7.67–15.03
20	error	0.492	0.495	0.488	0.000
	mean gap	0.13–0.23	0.11–0.16	0.13–0.23	6.74–10.90

5.2 Target and symmetry

Two classes of martingale differences, $x_t = \varepsilon_t \sqrt{1 + aS_t}$ with iid signs ε_t and an independent two-state Markov chain S_t, share the marginal law, the autocovariance and the conditional mean, and differ in how the conditional variance depends on the past. By §2, no readout of the original target has a mean signal, and a squared target creates one only together with a bias. Table 1 gives the four variants, each with an intercept and without per-sequence standardization. Only the readout with both remedies has a class-mean difference above 0.23 pooled standard deviations, and under shrinkage LDA only it separates the classes (Appendix F.3).

The same mechanisms appear on classes defined by families of stationary laws, with parameters drawn for each sequence (Appendix F.3). The bilinear and threshold classes are not symmetric under $x \mapsto -x$, and a bias alone brings the bilinear task from error 0.319 to 0.000. The AR and AR-GARCH classes share the law of their autoregressive coefficients and differ in conditional variance, and under LDA only the readout variant with both remedies leaves chance, with error 0.143.

6 Related work

Representing data by fitted parameters extends well beyond time series. Functa represent each image or shape by the modulation of a shared network fitted to it (Dupont et al., 2022), and hyper-representations and weight-space architectures learn from the weights of trained networks (Schürholt et al., 2022; Navon et al., 2023; Lim et al., 2024; Zhou et al., 2024). Qiu et al. (2026) note that a precise account of how data semantics are encoded in such weights is missing, and give one for implicit neural representations through the implicit function theorem. For time series, Norden et al. (2026) use structural identifiability to relate parameter values that produce one output, and Prosperino et al. (2026) use the distribution of fitted readout weights as a fingerprint of the input dynamics. Our analysis concerns the statistical side of this question, namely which properties of the generating law the fitted parameters retain and at what sampling cost. The closest statistical framework is distribution regression (Szabó et al., 2016), which learns from bags of iid samples embedded by kernel means and analyses both stages of sampling. Here the inner sample is one dependent trajectory, its embedding is a fitted predictor, and the size of the inner sample is the length L that enters the sampling law. Random convolutional kernels (Dempster et al., 2020) are a random-feature representation for time-series classification that pools features instead of fitting a predictor.

The theory of echo-state networks has mostly studied what a reservoir can do for one input law. Echo-state families approximate fading-memory filters densely (Grigoryeva & Ortega, 2018; Gonon & Ortega, 2020), the echo-state property (Jaeger, 2001; Yildiz et al., 2012) fixes the input-to-state map, ridge readouts fitted on one ergodic trajectory approximate the underlying dynamics (Hart et al., 2021), and recurrent kernel machines take the width of the reservoir to infinity (Hermans & Schrauwen, 2012). Information-processing and memory capacity (Dambre et al., 2012; Inubushi & Yoshimura, 2017; Ballarin et al., 2024), risk bounds (Gonon et al., 2020), random-matrix analyses (Couillet et al., 2016) and Volterra and kernel representations (Gonon et al., 2026) describe a reservoir under one input law. We study instead how the fitted readout differs between processes. The closest analysis is the unfolding of Nestler et al. (2020), which expands a rate network in the strength of a quadratic gain, so that stimulus cumulants enter the state's mean and covariance, and trains a readout across trials on the state at one time. Here the expansion is in the input scale, the nonlinearity is odd, so

even cumulants enter at odd orders, and the object is the readout of each sequence. For fourth-order lag moments, Proposition E.3 gives the geometric envelope of the memory motifs of Tiňo (2020).

Differences above second order are also the subject of classical time-series statistics. Higher-order spectra (Brillinger, 1965) and tests of time reversibility (Ramsey & Rothman, 1996) detect such differences once the order and the lags are chosen, and higher-order cumulants recover the phase of a linear filter, which second-order statistics do not determine (Lii & Rosenblatt, 1982). Path signatures (Király & Oberhauser, 2019) encode ordered higher-order lag structure and would also separate the reversal pairs. Next-generation reservoir computing (Gauthier et al., 2021; Grigoryeva et al., 2025) and kernels on nonlinear vector autoregressive embeddings (De Felice et al., 2023) use delay-polynomial features, of which the cubic term of the readout is a finite-rank image (Remark E.4). These statistics are built for a chosen order and chosen lags, whereas a fitted readout weights all orders and lags through its reservoir (§5.1).

7 Limitations

The graded orders and the sampling law are small-scale statements, proved for scalar input and, for the sampling law, for bounded moving averages at a fixed ridge. Proposition D.2 extends the cubic separation law, though not the lower bound, to vector input. At intermediate scales, the onset of detection is computed from exact moments in the Gaussian limit, on reservoirs of finite memory.

8 Conclusion

A fitted readout is a statistic of the law of the process, and what it can tell apart is set jointly by the target, the reservoir and the length of the trajectory. The target decides which differences reach the population readout, the order of a difference decides the power of the input scale at which it enters, and the linear noise of one trajectory sets the small-scale cost $\lambda^{4k}L$ of detecting it, with a further cost for estimating the geometry of the readout from few trajectories. When a difference stays invisible, the analysis points to what to change, namely the target, a bias or the scale, which is a parameter of the representation as much as of the reservoir. The restriction to the conditional mean and the sign symmetry hold for other fitted-parameter representations as well, and the same questions of visibility and sampling cost can be asked of them.

AI use statement

We used generative AI tools as assistants in developing the proofs, implementing the experiments and writing the paper. The research direction and the core ideas were set by the authors, who reviewed and verified all AI-assisted work. We take responsibility for the final content of this work.

Ethics statement

This paper analyses a representation and verifies the analysis numerically on synthetic stationary processes. We introduce no new data, and we see no ethical concern specific to this work.

Reproducibility statement

Every theorem, proposition and corollary is proved in full in Appendices A–E, with the assumptions each result needs stated in it. The experiments' witnesses, reservoirs, grids, classifiers and selection rules are specified in Appendix F. Every quoted number comes from a seeded run whose configuration is stated in the corresponding subsection. The code is in the supplementary material.

References

Witali Aswolinskiy, René Felix Reinhart, and Jochen J. Steil. Time series classification in reservoir- and model-space. *Neural Processing Letters*, 48(2):789–809, 2018. doi: 10.1007/s11063-017-9765-5.

Giovanni Ballarin, Lyudmila Grigoryeva, and Juan-Pablo Ortega. Memory of recurrent networks: Do we compute it right? *Journal of Machine Learning Research*, 25(243):1–38, 2024. URL https://jmlr.org/papers/v25/23-0568.html.

Filippo Maria Bianchi, Simone Scardapane, Sigurd Løkse, and Robert Jenssen. Reservoir computing approaches for representation and classification of multivariate time series. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2169–2179, 2021. doi: 10.1109/TNNLS.2020.3001377.

David R. Brillinger. An introduction to polyspectra. *The Annals of Mathematical Statistics*, 36(5):1351–1374, 1965. doi: 10.1214/aoms/1177699896.

Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer, New York, 2 edition, 1991.

Ao Chen, Xiren Zhou, Yizhan Fan, and Huanhuan Chen. Underground diagnosis based on GPR and learning in the model space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3832–3844, 2024. doi: 10.1109/TPAMI.2023.3347739.

Ao Chen, Xiren Zhou, and Huanhuan Chen. Efficient anomaly detection of irregular sequences in Ct-Echo model space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15731–15739, 2025. doi: 10.1609/aaai.v39i15.33727.

Huanhuan Chen, Peter Tiňo, Ali Rodan, and Xin Yao. Learning in the model space for cognitive fault diagnosis. *IEEE Transactions on Neural Networks and Learning Systems*, 25(1):124–136, 2014. doi: 10.1109/TNNLS.2013.2256797.

Huanhuan Chen, Fengzhen Tang, Peter Tiňo, Anthony G. Cohn, and Xin Yao. Model metric co-learning for time series classification. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-15)*, pp. 3387–3394. AAAI Press, 2015. URL https://www.ijcai.org/Abstract/15/477.

Romain Couillet, Gilles Wainrib, Hafiz Tiomoko Ali, and Harry Sevi. A random matrix approach to echo-state neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 517–525. PMLR, 2016. URL https://proceedings.mlr.press/v48/couillet16.html.

Joni Dambre, David Verstraeten, Benjamin Schrauwen, and Serge Massar. Information processing capacity of dynamical systems. *Scientific Reports*, 2:514, 2012. doi: 10.1038/srep00514.

Giovanni De Felice, John Y. Goulermas, and Vladimir V. Gusev. Time series kernels based on nonlinear vector autoregressive delay embeddings. In *Advances in Neural Information Processing Systems*, volume 36, pp. 37230–37251, 2023. doi: 10.52202/075280-1618.

Angus Dempster, François Petitjean, and Geoffrey I. Webb. ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5): 1454–1495, 2020. doi: 10.1007/s10618-020-00701-z.

Emilien Dupont, Hyunjik Kim, S. M. Ali Eslami, Danilo Jimenez Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you can treat it like one. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5694–5725, 2022.

Daniel J. Gauthier, Erik Bollt, Aaron Griffith, and Wendson A. S. Barbosa. Next generation reservoir computing. *Nature Communications*, 12:5564, 2021. doi: 10.1038/s41467-021-25801-2.

Lukas Gonon and Juan-Pablo Ortega. Reservoir computing universality with stochastic inputs. *IEEE Transactions on Neural Networks and Learning Systems*, 31(1):100–112, 2020. doi: 10.1109/TNNLS.2019.2899649.

Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega. Risk bounds for reservoir computing. *Journal of Machine Learning Research*, 21(240):1–61, 2020. URL https://www.jmlr.org/papers/v21/19-902.html.

Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega. Reservoir kernels and Volterra series. *IEEE Transactions on Neural Networks and Learning Systems*, 37(5):2181–2192, 2026. doi: 10.1109/TNNLS.2025.3630143.

Lyudmila Grigoryeva and Juan-Pablo Ortega. Echo state networks are universal. *Neural Networks*, 108:495–508, 2018. doi: 10.1016/j.neunet.2018.08.025.

Lyudmila Grigoryeva, Hannah Lim Jing Ting, and Juan-Pablo Ortega. Infinite-dimensional next-generation reservoir computing. *Physical Review E*, 111(3):035305, 2025. doi: 10.1103/PhysRevE.111.035305.

Allen G. Hart, James L. Hook, and Jonathan H. P. Dawes. Echo state networks trained by Tikhonov least squares are $L^2(\mu)$ approximators of ergodic dynamical systems. *Physica D: Nonlinear Phenomena*, 421:132882, 2021. doi: 10.1016/j.physd.2021.132882.

Michiel Hermans and Benjamin Schrauwen. Recurrent kernel machines: Computing with infinite echo state networks. *Neural Computation*, 24(1):104–133, 2012. doi: 10.1162/neco_a_00200.

Masanobu Inubushi and Kazuyuki Yoshimura. Reservoir computing beyond memory-nonlinearity trade-off. *Scientific Reports*, 7(1):10199, 2017. doi: 10.1038/s41598-017-10257-6.

Tommi Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems*, volume 11, pp. 487–493. MIT Press, 1999. URL https://papers.nips.cc/paper_files/paper/1998/hash/db1915052d15f7815c8b88e879465a1e-Abstract.html.

Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks. Technical Report GMD Report 148, German National Research Center for Information Technology, 2001.

Franz J. Király and Harald Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20(31):1–45, 2019. URL https://jmlr.org/papers/v20/16-314.html.

Steven G. Krantz and Harold R. Parks. *A Primer of Real Analytic Functions*. Birkhäuser, 2nd edition, 2002. doi: 10.1007/978-0-8176-8134-0.

Keh-Shin Lii and Murray Rosenblatt. Deconvolution and estimation of transfer function phase and coefficients for nongaussian linear processes. *The Annals of Statistics*, 10(4):1195–1208, 1982. doi: 10.1214/aos/1176345984.

Derek Lim, Haggai Maron, Marc T. Law, Jonathan Lorraine, and James Lucas. Graph metanetworks for processing diverse neural architectures. In *International Conference on Learning Representations (ICLR)*, pp. 11430–11458, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2feafd84b23a52b3fa434bc6d7682256-Abstract-Conference.html.

Shikang Liu, Chuyang Wei, Xiren Zhou, and Huanhuan Chen. Spectral-aware reservoir computing for fast and accurate time series classification. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39774–39788. PMLR, 2025. URL https://proceedings.mlr.press/v267/liu25by.html.

Shikang Liu, Ziyu Tang, Xiren Zhou, and Huanhuan Chen. SVGL: Scale-variable graph learning in model space for multivariate time series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23801–23809, 2026. doi: 10.1609/aaai.v40i28.39555.

Mantas Lukoševičius. A practical guide to applying echo state networks. In *Neural Networks: Tricks of the Trade*, volume 7700 of *Lecture Notes in Computer Science*, pp. 659–686. Springer, 2012. doi: 10.1007/978-3-642-35289-8_36.

Elizabeth Ann Maharaj. A significance test for classifying ARMA models. *Journal of Statistical Computation and Simulation*, 54(4):305–331, 1996. doi: 10.1080/00949659608811737.

Aviv Navon, Aviv Shamsian, Idan Achituve, Ethan Fetaya, Gal Chechik, and Haggai Maron. Equivariant architectures for learning in deep weight spaces. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25790–25816, 2023.

Sandra Nestler, Christian Keup, David Dahmen, Matthieu Gilson, Holger Rauhut, and Moritz Helias. Unfolding recurrence by Green's functions for optimized reservoir computing. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17380–17390, 2020.

Janis Norden, Elisa Oostwal, Michael Chappell, Peter Tiňo, and Kerstin Bunte. Structure is information: Structural identifiability mappings for machine learning with partially observed dynamical systems. *IEEE Transactions on Cybernetics*, 56(10):5423–5435, 2026. doi: 10.1109/TCYB.2026.3694145.

Frank W. J. Olver, Daniel W. Lozier, Ronald F. Boisvert, and Charles W. Clark (eds.). *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

Domenico Piccolo. A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*, 11(2):153–164, 1990. doi: 10.1111/j.1467-9892.1990.tb00048.x.

Davide Prosperino, Haochun Ma, and Christoph Räth. Distinguishing indistinguishable attractors: Unsupervised anomaly detection with reservoir computers. *arXiv preprint arXiv:2606.21322*, 2026.

Tianming Qiu, Christos Sonis, and Hao Shen. Ensuring semantics in weights of implicit neural representations through the implicit function theorem. *arXiv preprint arXiv:2601.23181*, 2026.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pp. 1177–1184, 2007. URL https://papers.nips.cc/paper_files/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

James B. Ramsey and Philip Rothman. Time irreversibility and business cycle asymmetry. *Journal of Money, Credit and Banking*, 28(1):1–21, 1996. doi: 10.2307/2077963.

R. Ranga Rao. Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics*, 33(2):659–680, 1962. doi: 10.1214/aoms/1177704588.

Konstantin Schürholt, Boris Knyazev, Xavier Giró-i Nieto, and Damian Borth. Hyper-representations as generative models: Sampling unseen neural network weights. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27906–27920, 2022. doi: 10.52202/068431-2024.

Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016.

Peter Tiňo. Dynamical systems as temporal feature spaces. *Journal of Machine Learning Research*, 21(44): 1–42, 2020.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998. doi: 10.1017/CBO9780511802256.

Gideon Weiss. Time-reversibility of linear stochastic processes. *Journal of Applied Probability*, 12(4):831–836, 1975. doi: 10.2307/3212735.

Wei Biao Wu. Nonlinear system theory: another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154, 2005. doi: 10.1073/pnas.0506715102.

Izzet B. Yildiz, Herbert Jaeger, and Stefan J. Kiebel. Re-visiting the echo state property. *Neural Networks*, 35: 1–9, 2012. doi: 10.1016/j.neunet.2012.07.005.

Allan Zhou, Chelsea Finn, and James Harrison. Universal neural functionals. In *Advances in Neural Information Processing Systems*, volume 37, pp. 104754–104775, 2024. doi: 10.52202/079017-3326.

Xiren Zhou, Ziyu Tang, Shikang Liu, Ao Chen, Xiangyu Wang, and Huanhuan Chen. Fault diagnosis in REDNet model space. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, pp. 7209–7217, 2025. doi: 10.24963/ijcai.2025/802.

Xiren Zhou, Chuyang Wei, Ao Chen, Shikang Liu, Xiangyu Wang, and Huanhuan Chen. Fault diagnosis of irregular sequences by adjoint learning in continuous-time model space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29044–29052, 2026. doi: 10.1609/aaai.v40i34.40141.

Guide to the appendix. Appendix A fixes the notation. Appendices B to E prove the results, in the order of the body, and Appendix F gives the protocols and the full results of the experiments. Table 2 lists where each result of the body is proved and measured. Each long proof opens with a short paragraph on its idea.

Table 2: Where each result of the body is proved and measured.

Result in the body	Proof	Experiments
Raw sequences against fitted readouts (Proposition B.1)	Appendix B	Appendix F.2
Graded orders (Theorem 3)	Appendix C.2.1	Appendix F.4
Memory and the scale axis (Propositions D.6, E.3, D.8, Corollary D.9)	Appendices D.4, E.2	Appendices F.7, F.8, F.10
Blind spots and remedies (Remark 2, Corollary 1)	§2, Proposition C.4, Remark E.2	Appendix F.3
Sampling law and lower bound (Proposition D.2, Theorem 4)	Appendices D.2, D.5, D.6	Appendices F.5, F.6
Fixed-order statistics	–	Appendices F.9, F.11

A Notation and the population readout

This section collects the notation, fixes the convergence convention used in every proof below, and proves that the fitted readout converges to a population readout whose class difference splits into a moment term and a covariance term.

Table 3: Summary of notation.

Symbol	**Meaning**
n, d_x, T, y	Reservoir width, input channels, sequence length, class index
t_{w}, s, $I_s = \{t_{\mathrm{w}}+1, \dots, T-s\}$, $L = \lvert I_s\rvert$	Washout, target span, fit window, window length
W, W_{in}, $\rho(W)$, q	Reservoir and input weights (fixed), spectral radius, contraction constant
λ, γ, g	Input scale, fixed ridge, relative ridge level $\gamma = g\,\mathrm{tr}(\hat{\Sigma}_h)/n$
h_t; $h_t^{(1)} = a_t$, $h_t^{(3)} = c_t$, b_t	$\tanh$ state; linearized, cubic and scaled-residual states of (E.2)
$\hat{\Sigma}_h$, $\hat{M}$, $\hat{W}$	Sample state covariance, target–state cross-moment, ridge readout (2)
$\hat{M}_1$, $\hat{\Sigma}_{11}$; $\hat{M}_{3,\lambda}$, $\hat{M}_{3,0}$	Linear block and its covariance; cubic block, cubic statistic
$M_{g,P}$, $\Sigma_{h,P}$, $\bar{W}_{g,P}$; $M_{g,P}^{(2k+1)}$	Population moments and readout; coefficient of λ^{2k+1} in the cross-moment
$\Gamma_P(\tau)$	Autocovariance
$c_k(s)$	Witness $\Delta\,\mathrm{cum}(x_{t+s}, x_t, \dots, x_t)$ with $2k+1$ copies of x_t
δ_k, κ_k, $\kappa_\star = \kappa_1$, $\langle \hat{W} \rangle_L$	Signal and separation constant of Theorem 4, mean readout over the two classes
ϑ_j, e_t, B_e, B_x	Moving-average coefficients, innovations, their bound, the input bound
p, ϖ_t, F_λ, G_λ, $(\hat{\Lambda}, \hat{Y})$	Memory, window, frame, Gram matrix and window statistic of Proposition D.6
$\Delta A := A_1 - A_2$	Class difference of any class-indexed quantity
V_1, $\bar{V}_1$, $R_\lambda = (\gamma I + \lambda^2 \Sigma_{11})^{-1}$, d^2	Long-run covariance of $\sqrt{L}\hat{M}_1$, pooled, population resolvent, Mahalanobis separation of class means

Convergence convention. Throughout, $\rho(W) < 1$ and Gelfand's formula fix $\xi \in (\rho(W), 1)$ and

$$C_W := \sup_{k \geq 0} \|W^k\|_{\mathrm{op}} \xi^{-k} < \infty, \tag{A.1}$$

so $\|W^k\|_{\mathrm{op}} \le C_W \xi^k$. For normal W one can take $C_W = 1$ and $\xi = \rho(W)$, and otherwise C_W accounts for the transient amplification.

Every Neumann series $\sum_{k\ge 0} W^k v_{t-k}$ below has a stationary summand of some fixed polynomial degree in the input path, and the hypothesis in force guarantees the required moment of this summand. The geometric bound then makes such a series converge absolutely almost surely and in $L^2(P)$. It also makes the stationary causal solution unique, because the homogeneous recursion forces $u_t = W^K u_{t-K} \to 0$, and it justifies interchanging expectation and summation.

A.1 Population identity

Proposition A.1 (Population limit of the fitted readout)**.** *Let Assumption 2 hold, so that the stationary state h_t exists (Proposition C.4), and let the input be stationary and ergodic with finite second moments. With $\Sigma_{h,y} = \mathbb{E}_y[h_t h_t^\top]$ and $M_{g,y} = \mathbb{E}_y[x_{t+s} h_t^\top]$, the fitted readout of a sequence of class y converges as $L \to \infty$ to*

$$\bar{W}_{g,y} := M_{g,y}(\Sigma_{h,y} + \gamma I)^{-1}, \tag{A.2}$$

and

$$\Delta\bar{W}_g \;=\; \Delta M_g \,(\Sigma_{h,1} + \gamma I)^{-1} \;-\; M_{g,2}\,(\Sigma_{h,1} + \gamma I)^{-1}\,\Delta\Sigma_h\,(\Sigma_{h,2} + \gamma I)^{-1}. \tag{A.3}$$

Under Assumption 2 the trajectory from $h_0 = 0$ converges to this state (Proposition C.4).

Proof of Proposition A.1. Since (x_{t+s}, h_t) is stationary and ergodic, Birkhoff's theorem gives $\hat{\Sigma}_{h,y} \to \Sigma_{h,y}$ and $\hat{M}_{g,y} \to M_{g,y}$ almost surely. Since $(A, B) \mapsto A(B + \gamma I)^{-1}$ is continuous on $B \succeq 0$ for $\gamma > 0$, it follows that $\hat{W}_{g,y} \to \bar{W}_{g,y}$ almost surely.

For the decomposition, we write

$$\Delta\bar{W}_g = \Delta M_g(\Sigma_{h,1} + \gamma I)^{-1} + M_{g,2}\big[(\Sigma_{h,1} + \gamma I)^{-1} - (\Sigma_{h,2} + \gamma I)^{-1}\big], \tag{A.4}$$

and the resolvent identity

$$(A + \gamma I)^{-1} - (B + \gamma I)^{-1} = -(A + \gamma I)^{-1}(A - B)(B + \gamma I)^{-1} \tag{A.5}$$

gives (A.3). □

B Proof of Proposition B.1

Proposition B.1 (Raw sequences and fitted readouts)**.** *Let $d_x = 1$, and let each class be a process x_t with zero mean and unit variance that is a moving average of finite order of iid symmetric innovations with finite fourth moment. Fix $N \ge 1$ training sequences per class and an independent test sequence, all of length L.*

(a) Raw sequences. *For every odd $k \le 2N$, the error of the k-nearest-neighbour rule in the Euclidean distance on $\mathbb{R}^L$, under equal class probabilities, converges as $L \to \infty$ to a limit $e_k > 0$. The limit depends on the classes only through*

$$c_{ab} := \sum_h \Gamma_a(h)\Gamma_b(h) \tag{B.1}$$

and the long-run variances v_b of x_t^2, and $e_k = \frac{1}{2}$ whenever $\Gamma_1 = \Gamma_2$.

(b) Fitted readouts. *Let Assumption 2 hold and $\bar{W}_1 \ne \bar{W}_2$. For every $N \ge 1$, the error of the nearest-neighbour rule on the fitted readouts, in any norm, tends to zero as $L \to \infty$.*

Idea. For part (a), the Euclidean rule on raw sequences is reduced to a comparison of centred scores Z_i, one per training sequence. A conditional central limit theorem shows that these scores become independent Gaussians whose variances depend on the classes only through v_b and c_{ab}, so the error converges to a positive limit. For part (b), every fitted readout converges to the population readout of its class, and the nearest neighbour then has the right class.

Proof. (a) Write

$$S_u := \sum_t u_t^2, \qquad C_{uv} := \sum_t u_t v_t, \tag{B.2}$$

so that $\|x-y\|^2 = S_x + S_y - 2C_{xy}$. The rule ranks the training sequences y_i by $\|x-y_i\|^2$, equivalently by

$$Z_i := (S_{y_i} - L - 2C_{xy_i})/\sqrt{L}, \tag{B.3}$$

since S_x is common to all of them. Let the test sequence x be of class a and condition on it. The pairs (S_{y_i}, C_{xy_i}) are then independent across i.

For y_i of class b, the vector $(S_{y_i} - L,\ C_{xy_i})$ is the sum over t of $(y_{i,t}^2 - 1,\ x_t y_{i,t})$, a finitely dependent array given x. Its conditional covariance divided by L converges almost surely to $\mathrm{diag}(v_b, c_{ab})$. Indeed, the conditional variance of C_{xy_i} is $\sum_{t,t'} x_t x_{t'} \Gamma_b(t-t')$, and by the ergodic theorem

$$\frac{1}{L}\sum_t x_t x_{t+h} \to \Gamma_a(h) \tag{B.4}$$

for each of the finitely many lags h with $\Gamma_b(h) \neq 0$. The conditional covariance of S_{y_i} and C_{xy_i} is $\sum_{t,t'} x_{t'} \mathbb{E}[y_{i,t}^2 y_{i,t'}] = 0$, because the innovations are symmetric.

Since $\max_{t \le L} x_t^2 = o(L)$ almost surely, the Lindeberg condition holds, and the central limit theorem for finitely dependent arrays gives, conditionally on x and almost surely, that the Z_i converge jointly to independent $\mathcal{N}(0, \sigma_{ab}^2)$ variables with $\sigma_{ab}^2 := v_b + 4c_{ab}$.

The limit does not depend on x, so the convergence also holds unconditionally. Ties have probability zero under the limit law, so the set of indices of the k smallest Z_i converges in law, and the error of the rule converges to the error e_k of the same rule applied to independent Gaussian scores with these variances. Every set of k indices has positive probability of being the k smallest under this law, so $e_k > 0$. If $\Gamma_1 = \Gamma_2$, then c_{ab} does not depend on a, so the law of the scores, and of the output of the rule, is the same for both test classes. The error is then $\frac{1}{2}$.

(b) Under Assumption 2 the fitted readout converges almost surely to the population readout of its class (Proposition C.4). The same holds at the relative ridge $g\,\mathrm{tr}(\hat{\Sigma}_h)/n$, since $\mathrm{tr}\,\hat{\Sigma}_h \to \mathrm{tr}\,\Sigma_h > 0$ and the readout is continuous in the ridge. Let $\delta := \|\bar{W}_1 - \bar{W}_2\| > 0$. Almost surely, for L large enough the $2N+1$ fitted readouts lie within $\delta/3$ of their population readouts, and the training readout nearest to the test readout has its class, so the error tends to zero by dominated convergence. □

In the limit of part (a), the squared distance of the test sequence to a training sequence of class b is $2L + \sqrt{L}(Z_x + Z_i)$ with Z_x common to all training sequences, so the rule compares centred Gaussian scores whose variances depend on the classes only through v_b and c_{ab}. At $N = 1$ and $k = 1$ the limit is $\frac{1}{2}$ for every pair of classes, since the difference of two independent centred Gaussian scores is symmetric. For $N \ge 2$ and $\Gamma_1 \neq \Gamma_2$ the limit is below $\frac{1}{2}$ in the cases we computed (Appendix F.2), but close to it. On the moving averages of Appendix F.2 it is 0.496 at $N = 5$ and 0.492 at $N = 25$.

C Proofs for the second and the higher orders

This section proves Proposition C.1 on the linearized readout and Theorem 3 on the order in λ at which a difference of moments enters the $\tanh$ readout.

C.1 Second order

At first order in λ the readout depends on the input law only through the autocovariance, and at finite length, time reversal changes it by $O(1/T)$.

Proposition C.1 (Linearized moments are determined by the autocovariance)**.** *Under Assumption 1, for every stationary P with finite second moments:*

(a) *The stationary moments of the linearized state are*

$$M^{\mathrm{lin}}_{g_s,P} = \sum_{k\ge 0} \Gamma_P(s+k) W_{\mathrm{in}}^\top (W^\top)^k, \qquad \Sigma^{\mathrm{lin}}_{h,P} = \sum_{j,k\ge 0} W^j W_{\mathrm{in}} \Gamma_P(k-j) W_{\mathrm{in}}^\top (W^\top)^k, \tag{C.1}$$

so $\bar{W}^{\mathrm{lin}}_{g_s,P}$ depends on P only through $\{\Gamma_P(\tau)\}$.

(b) *If $\Gamma_{P_1} = \Gamma_{P_2}$ then $\bar{W}^{\mathrm{lin}}_{g_s,P_1} = \bar{W}^{\mathrm{lin}}_{g_s,P_2}$ for every $s \ge 0$ and every $\gamma > 0$. Proof in Appendix C.1.1.*

C.1.1 Proof of Proposition C.1 (The linearized moments are the autocovariance)

Proof. **Part (a): Autocovariance factorization.** Iterating the linear recurrence $h_t^{\mathrm{lin}} = W h_{t-1}^{\mathrm{lin}} + W_{\mathrm{in}} x_t$ with zero initial condition at $-\infty$ gives the convergent series

$$h_t^{\mathrm{lin}} = \sum_{k=0}^{\infty} W^k W_{\mathrm{in}} x_{t-k}. \tag{C.2}$$

Taking expectations,

$$\begin{aligned} M_{g_s,P}^{\mathrm{lin}} &= \mathbb{E}_P\Big[x_{t+s} \sum_{k=0}^{\infty} x_{t-k}^\top W_{\mathrm{in}}^\top (W^\top)^k\Big] \\ &= \sum_{k=0}^{\infty} \mathbb{E}_P[x_{t+s} x_{t-k}^\top]\, W_{\mathrm{in}}^\top (W^\top)^k = \sum_{k=0}^{\infty} \Gamma_P(s+k)\, W_{\mathrm{in}}^\top (W^\top)^k, \end{aligned} \tag{C.3}$$

where we used stationarity, $\mathbb{E}[x_{t+s} x_{t-k}^\top] = \Gamma_P(s+k)$. For the state covariance,

$$\begin{aligned} \Sigma_{h,P}^{\mathrm{lin}} &= \mathbb{E}_P\Big[\Big(\sum_{j=0}^{\infty} W^j W_{\mathrm{in}} x_{t-j}\Big)\Big(\sum_{k=0}^{\infty} W^k W_{\mathrm{in}} x_{t-k}\Big)^\top\Big] \\ &= \sum_{j,k=0}^{\infty} W^j W_{\mathrm{in}}\, \mathbb{E}_P[x_{t-j} x_{t-k}^\top]\, W_{\mathrm{in}}^\top (W^\top)^k = \sum_{j,k=0}^{\infty} W^j W_{\mathrm{in}}\, \Gamma_P(k-j)\, W_{\mathrm{in}}^\top (W^\top)^k. \end{aligned} \tag{C.4}$$

Both expressions depend on P only through the autocovariance sequence $\{\Gamma_P(\tau)\}_{\tau\in\mathbb{Z}}$, proving part (a) of Proposition C.1. The ridge readout $\bar{W}_{g_s,P}^{\mathrm{lin}} = M_{g_s,P}^{\mathrm{lin}}(\Sigma_{h,P}^{\mathrm{lin}} + \gamma I)^{-1}$ is a continuous function of the two moment matrices, and hence it depends on the input law only through the same sequence.

Part (b): Null space. If $\Gamma_{P_1} = \Gamma_{P_2}$, then by part (a) $M_{g_s,P_1}^{\mathrm{lin}} = M_{g_s,P_2}^{\mathrm{lin}}$ and $\Sigma_{h,P_1}^{\mathrm{lin}} = \Sigma_{h,P_2}^{\mathrm{lin}}$, so $\bar{W}_{g_s,P_1}^{\mathrm{lin}} = \bar{W}_{g_s,P_2}^{\mathrm{lin}}$ immediately. □

Lemma C.2 (Time reversal at finite length)**.** *Let $d_x = 1$, let P be stationary with $\mathbb{E}x_t^4 < \infty$, and let $\tilde{P}$ be its time reversal, that is, the law of $(x_{-t})_{t\in\mathbb{Z}}$. Consider the linearized recurrence with $h_0 = 0$ and $\|W^k\| \le Cq^k$, $q < 1$. For any span s, ridge $\gamma > 0$ and the fit window I_s of §2, let $\hat{W}(x)$ be the readout of the path $x = (x_1, \dots, x_T)$, and let $x^R = (x_T, \dots, x_1)$.*

Then $\mathbb{E}_P\|\hat{W}(x^R) - \hat{W}(x)\| \le C'/T$, where C' depends on $C, q, \gamma, s, t_{\mathrm{w}}$ and the fourth moment. Moreover, $\hat{W}$ under $\tilde{P}$ has the law of $\hat{W}(x^R)$ under P. The readout laws of P and $\tilde{P}$ therefore coincide up to an L^1 perturbation of order $1/T$.

Proof. Write $h_t = \sum_{k<t} W^k W_{\mathrm{in}} x_{t-k}$. Both fitted moments are linear in the lagged sums

$$Q_{jk}(x) := \sum_{t=t_{\mathrm{w}}}^{T-1} x_{t-j} x_{t-k}, \qquad j, k \ge -s, \tag{C.5}$$

with weights bounded by $C^2 q^{j+k}$:

$$\hat{\Sigma}_h = \frac{1}{T} \sum_{j,k\ge 0} W^j W_{\mathrm{in}}\, Q_{jk}\, W_{\mathrm{in}}^\top (W^\top)^k, \qquad \hat{M} = \frac{1}{T} \sum_{k\ge 0} Q_{-s,k}\, (W^k W_{\mathrm{in}})^\top, \tag{C.6}$$

where the indices are restricted to the path. On the reversed path, $Q_{jk}(x^R)$ is a sum of the same lag-$|j-k|$ products $x_a x_{a+j-k}$ over a window of the same length. This window is shifted by at most $j + k + 2s + t_{\mathrm{w}}$ positions, so

$$\mathbb{E}\big|Q_{jk}(x^R) - Q_{jk}(x)\big| \;\le\; 2\,(j + k + 2s + t_{\mathrm{w}} + 2)\,\Gamma(0), \tag{C.7}$$

and summing the geometric weights gives

$$\mathbb{E}\|\Delta\hat{\Sigma}_h\| + \mathbb{E}\|\Delta\hat{M}\| = O(1/T), \tag{C.8}$$

with a constant that depends on C, q, s and t_{w}. The readout $\hat{M}(\hat{\Sigma}_h + \gamma I)^{-1}$ is Lipschitz in $(\hat{M}, \hat{\Sigma}_h)$ with constant $\gamma^{-1}(1 + \|\hat{M}\|/\gamma)$, and the Cauchy–Schwarz inequality bounds $\mathbb{E}\|\hat{M}\| \, \|\Delta\hat{\Sigma}_h\|$ in terms of the fourth moment. This gives the L^1 bound, and stationarity gives the equality in law of $(x_1, \dots, x_T)$ under $\tilde{P}$ and of x^R under P. □

The perturbation is small compared with the sampling fluctuations of the fitted moments, which are of order $1/\sqrt{T}$ where these moments obey a central limit theorem, as they do for the linear processes used here. In particular $\sqrt{T}$ times the difference of the linearized cross-moments tends to zero in L^1, so the cross-moments of a process and of its reversal have one limiting law and share V_1.

C.2 Statistical order

This subsection proves Theorem 3. The theorem describes the stationary solution of the $\tanh$ recurrence and its expansion in the input scale, and its proof rests on three facts, which we state and prove first.

1. Under Assumption 2 the recurrence contracts pathwise, so a unique stationary branch exists, and the pipeline computes this branch from $h_0 = 0$ (Lemma C.3, Proposition C.4).
2. The formal power series of the branch in λ has a unique stationary coefficient at every order, and the even orders vanish (Lemma C.5).
3. The branch depends smoothly on λ on an L^p scale, its Taylor coefficients at $\lambda = 0$ are the formal ones, and the remainder after order $2K + 1$ is $O(\lambda^{2K+3})$ (Lemma C.6).

The first and the third fact use Assumption 2, and the second needs only $\rho(W) < 1$.

Lemma C.3 (One-step contraction in the D-norm)**.** *Under Assumption 2, for any $v, v' \in \mathbb{R}^n$ and any $u \in \mathbb{R}^n$,*

$$\|\tanh(Wv + u) - \tanh(Wv' + u)\|_D \;\le\; q\,\|v - v'\|_D, \qquad \|z\|_D := (z^\top D z)^{1/2}, \tag{C.9}$$

where $\tanh$ *acts componentwise. The bound is uniform in the drive u.*

Proof. Scalar $\tanh$ is 1-Lipschitz and acts componentwise, so $|\tanh(a)_i - \tanh(b)_i| \le |a_i - b_i|$ for every coordinate i. Since D is diagonal with positive entries, $\|\cdot\|_D^2$ is a positively weighted sum of squared coordinates and therefore preserves this componentwise domination:

$$\begin{aligned} \|\tanh(a) - \tanh(b)\|_D^2 &= \sum_i D_{ii}\,|\tanh(a)_i - \tanh(b)_i|^2 \\ &\le \sum_i D_{ii}\,|a_i - b_i|^2 = \|a - b\|_D^2. \end{aligned} \tag{C.10}$$

We apply this with $a = Wv + u$ and $b = Wv' + u$, so that the drive u cancels, and then bound

$$\|W(v - v')\|_D^2 = (v - v')^\top W^\top D W (v - v') \le q^2 \|v - v'\|_D^2 \tag{C.11}$$

by Assumption 2. □

Proposition C.4 (Contractive realization of the nonlinear readout)**.** *Let Assumption 2 hold and let $(x_t)_{t\in\mathbb{Z}}$ be strictly stationary and ergodic. Then for every $\lambda \in \mathbb{R}$:*

1. *the recurrence $h_t = \tanh(W h_{t-1} + \lambda W_{\mathrm{in}} x_t)$ admits a unique bounded stationary causal solution $(h_t^{\star,\lambda})$,*
2. *for the trajectory $(\tilde{h}_t^{\lambda})$ generated from any deterministic initial state on the same input path,*

$$\|\tilde{h}_t^{\lambda} - h_t^{\star,\lambda}\|_D \;\le\; q^t\,\|\tilde{h}_0^{\lambda} - h_0^{\star,\lambda}\|_D \qquad \textit{almost surely}, \tag{C.12}$$

3. *if the target (y_t) is, jointly with (x_t), a stationary measurable function of the input process and $\mathbb{E}\|y_0\| < \infty$, then the empirical moments of* (2) *formed along $(\tilde{h}_t^{\lambda})$ converge almost surely as $L \to \infty$ to the stationary population moments*

$$M_{g,P}^{\lambda} := \mathbb{E}_P\big[y_0\,(h_0^{\star,\lambda})^\top\big], \qquad \Sigma_{h,P}^{\lambda} := \mathbb{E}_P\big[h_0^{\star,\lambda}(h_0^{\star,\lambda})^\top\big], \tag{C.13}$$

4. *consequently, for every $\gamma > 0$ the ridge readout* (2) *fitted from the $h_0 = 0$ trajectory converges almost surely to*

$$\bar{W}_{g,P}^{\tanh;\lambda} = M_{g,P}^{\lambda}\big(\Sigma_{h,P}^{\lambda} + \gamma I\big)^{-1}. \tag{C.14}$$

No boundedness of the input is used, so the unbounded Gaussian and Laplace witnesses are covered. The only moment condition is $\mathbb{E}\|y_0\| < \infty$ in part (3), which for $g_s(x)_t = x_{t+s}$ becomes $\mathbb{E}\|x_0\| < \infty$.

Idea. Every part follows from the one-step contraction of Lemma C.3, applied path by path. Starting the recurrence further and further in the past gives a Cauchy sequence whose limit is the stationary branch, and the same contraction makes any two trajectories on one input path merge at rate q. The effect of the initialization on the empirical moments is then an average of a geometric sequence, Birkhoff's theorem handles the stationary part, and continuity of the ridge map carries the limits over to the readout.

Proof. Fix λ and abbreviate $h_t^\star = h_t^{\star,\lambda}$, $\tilde{h}_t = \tilde{h}_t^\lambda$, $u_t = \lambda W_{\mathrm{in}} x_t$. Put $d_{\min} := \min_i D_{ii} > 0$ and $d_{\max} := \max_i D_{ii}$, so that

$$\sqrt{d_{\min}}\,\|v\|_2 \;\le\; \|v\|_D \;\le\; \sqrt{d_{\max}}\,\|v\|_2 \qquad \text{for all } v \in \mathbb{R}^n : \tag{C.15}$$

that is, $\|\cdot\|_D$ and $\|\cdot\|_2$ are equivalent, with constants depending only on D. All estimates below are made path by path, on a fixed realization of (x_t).

(1) Existence and uniqueness. Let $h_t^{[-m]}$ be the state at time t obtained by setting the state at time $t-m$ to 0 and iterating forward along the given input path. Consecutive pull-backs share their inputs from time $t-m$ onwards. Since both states lie in $[-1,1]^n$ after one update, they differ at that time by at most $(\mathrm{tr}D)^{1/2}$. Hence m applications of Lemma C.3 give

$$\|h_t^{[-m-1]} - h_t^{[-m]}\|_D \le q^m (\mathrm{tr}D)^{1/2}, \tag{C.16}$$

and the sequence is Cauchy on every input path.

Its limit $h_t^\star$ is a measurable function of the input past and takes values in $[-1,1]^n$. By continuity of $\tanh$ it solves the recurrence. It is given by a map that does not depend on t, so it commutes with the time shift, and $(x_t, h_t^\star)$ is jointly stationary and ergodic.

Two bounded stationary causal solutions start within $2(\mathrm{tr}D)^{1/2}$ of each other. Hence they differ by at most $2q^m(\mathrm{tr}D)^{1/2}$ for every m, and so they coincide almost surely.

(2) Geometric forgetting. The two trajectories are driven by the same inputs, so t applications of Lemma C.3 give the stated bound deterministically on each path, and hence almost surely. Since $h_0^\star \in [-1,1]^n$ and $\tilde{h}_0$ is deterministic, the initial gap is bounded by the path-independent constant $c_0 := \|\tilde{h}_0\|_D + (\mathrm{tr}D)^{1/2}$. For the implemented initialization $\tilde{h}_0 = 0$, we have $c_0 = (\mathrm{tr}D)^{1/2}$.

(3) Empirical moments. Let T denote the number of averaged terms, indexed $t = 1, \dots, T$. This is the window length L of (2). The washout only makes the initialization terms below smaller, so we may ignore it.

We first replace $\tilde{h}_t$ by $h_t^\star$ in both empirical moments. Using $\|aa^\top - cc^\top\|_F \le (\|a\|_2 + \|c\|_2)\|a - c\|_2$, the bound $\|\tilde{h}_t\|_2 \vee \|h_t^\star\|_2 \le \sqrt{n}$ valid for $t \ge 1$, part (2), and the left inequality in (C.15),

$$\frac{1}{T}\sum_{t=1}^{T} \|\tilde{h}_t\tilde{h}_t^\top - h_t^\star (h_t^\star)^\top\|_F \;\le\; \frac{2\sqrt{n}\,c_0}{\sqrt{d_{\min}}} \cdot \frac{1}{T}\sum_{t=1}^{T} q^t \;\le\; \frac{2\sqrt{n}\,c_0}{\sqrt{d_{\min}}\,(1-q)} \cdot \frac{1}{T} \;\longrightarrow\; 0 \tag{C.17}$$

deterministically. Similarly,

$$\frac{1}{T}\sum_{t=1}^{T} \|y_t(\tilde{h}_t - h_t^\star)^\top\|_F \;\le\; \frac{c_0}{\sqrt{d_{\min}}} \cdot \frac{1}{T}\sum_{t=1}^{T} q^t \|y_t\|_2, \tag{C.18}$$

and since (y_t) is stationary with $\mathbb{E}\|y_0\| < \infty$ we have

$$\mathbb{E}\sum_{t\ge 1} q^t \|y_t\| = \frac{q}{1-q}\mathbb{E}\|y_0\| < \infty, \tag{C.19}$$

so $\sum_{t\ge1} q^t\|y_t\| < \infty$ almost surely, and the Cesàro average above tends to 0 almost surely.

The equivalence (C.15) is used at this step. Part (2) controls the D-norm, whereas the moments are Euclidean, and passing from one norm to the other introduces only the fixed deterministic factor $d_{\min}^{-1/2}$, which cannot affect almost-sure convergence.

The process $(y_t, h_t^\star)$ is a measurable factor of the stationary ergodic input process, so it is stationary and ergodic. It is also integrable, since $\|h_0^\star(h_0^\star)^\top\|_F \le n$ and $\mathbb{E}\|y_0(h_0^\star)^\top\|_F \le \sqrt{n}\,\mathbb{E}\|y_0\| < \infty$. Birkhoff's ergodic theorem therefore gives

$$\frac{1}{T}\sum_t h_t^\star(h_t^\star)^\top \to \Sigma_{h,P}^\lambda \qquad \text{and} \qquad \frac{1}{T}\sum_t y_t(h_t^\star)^\top \to M_{g,P}^\lambda \tag{C.20}$$

almost surely. Adding the two displays transfers both limits to the initialized trajectory, which is the claim.

(4) Ridge readout. Both $\hat{\Sigma}_h$ and $\Sigma_{h,P}^\lambda$ are positive semidefinite, so $\|(\cdot+\gamma I)^{-1}\|_{\mathrm{op}} \le \gamma^{-1}$ uniformly in T and at the limit. The resolvent identity

$$(B+\gamma I)^{-1} - (B'+\gamma I)^{-1} = (B+\gamma I)^{-1}(B'-B)(B'+\gamma I)^{-1} \tag{C.21}$$

bounds the difference of two such resolvents by $\gamma^{-2}\|B-B'\|_{\mathrm{op}}$. Hence $(A,B) \mapsto A(B+\gamma I)^{-1}$ is continuous on $\mathbb{R}^{d_x\times n} \times \{B \succeq 0\}$ for each fixed $\gamma > 0$. Applying this map to the almost-sure limits of part (3) gives

$$\hat{M}_g(\hat{\Sigma}_h+\gamma I)^{-1} \to M_{g,P}^\lambda(\Sigma_{h,P}^\lambda+\gamma I)^{-1} \tag{C.22}$$

almost surely. □

The formal expansion. Substituting a power series in λ into the recurrence and matching orders gives a recursion for the coefficients. This recursion uses only the condition $\rho(W) < 1$.

Lemma C.5 (The formal coefficients)**.** *Let $\rho(W) < 1$ and let (x_t) be stationary with finite moments of every order. Write*

$$\tanh(u) = \sum_{j\ge 0} \tau_{2j+1} u^{\odot(2j+1)}, \qquad \tau_1 = 1, \quad \tau_3 = -\frac{1}{3}, \quad \tau_5 = \frac{2}{15}, \quad \tau_7 = -\frac{17}{315}, \ \ldots, \tag{C.23}$$

and substitute

$$h_t(\lambda) = \sum_{p\ge 1} \lambda^p h_t^{(p)} \tag{C.24}$$

into $h_t = \tanh(Wh_{t-1} + \lambda W_{\mathrm{in}} x_t)$. Matching powers of λ gives, for every $k \ge 0$,

$$h_t^{(2k+1)} = W h_{t-1}^{(2k+1)} + F_{2k+1}\big(h^{(1)}, h^{(3)}, \ldots, h^{(2k-1)}\big)_t, \tag{C.25}$$

where $F_1 = W_{\mathrm{in}} x_t$ and, for $k \ge 1$, F_{2k+1} is a polynomial in the lower-order coefficients of weighted degree $2k+1$. Here the coefficient $h^{(2j+1)}$ has weight $2j+1$, and the coefficients of the polynomial are determined by $\{\tau_{2j+1}\}_{j\le k}$ and W. The even-order coefficients are identically zero. Each recursion has exactly one stationary $L^2(P)$ solution, namely the Neumann series

$$h_t^{(2k+1)} = \sum_{m\ge 0} W^m F_{2k+1}\big(h^{(1)}, \ldots, h^{(2k-1)}\big)_{t-m}, \tag{C.26}$$

and $h^{(2k+1)}$ is a homogeneous polynomial of degree $2k+1$ in the input path. In particular

$$h_t^{(1)} = \sum_{k\ge 0} W^k W_{\mathrm{in}} x_{t-k}, \qquad h_t^{(3)} = -\tfrac{1}{3}\sum_{k\ge 0} W^k \big(h_{t-k}^{(1)}\big)^{\odot 3}: \tag{C.27}$$

the first-order coefficient is the linearized state of Proposition C.1, and the cubic one accumulates $-\frac{1}{3}$ times the elementwise cubes of the linearized state through the reservoir recurrence.

Proof. At odd power $2k+1$ the matching gives

$$\begin{aligned} h_t^{(2k+1)} &= [\lambda^{2k+1}]\big(Wh_{t-1}(\lambda) + \lambda W_{\mathrm{in}} x_t\big) \\ &\quad + \sum_{j=1}^{k} \tau_{2j+1}\,[\lambda^{2k+1}]\big(Wh_{t-1}(\lambda) + \lambda W_{\mathrm{in}} x_t\big)^{\odot(2j+1)}. \end{aligned} \tag{C.28}$$

The first term is $Wh_{t-1}^{(2k+1)}$, with $W_{\mathrm{in}} x_t$ contributing only at $k = 0$. Each $(2j+1)$-fold product in the second term, expanded over index tuples $(p_1, \ldots, p_{2j+1})$ with $\sum_m p_m = 2k+1$, involves

only coefficients $h^{(p_m)}$ with $p_m \le 2k-1$, since $j \ge 1$. Collecting these terms gives F_{2k+1}, and the weight count is the constraint $\sum_m p_m = 2k+1$.

Only odd Taylor coefficients of $\tanh$ are non-zero, so the matching at even power $2k$ gives $h_t^{(2k)} = Wh_{t-1}^{(2k)}$, whose only square-integrable stationary solution is zero. At $k=1$ the recursion is

$$h_t^{(3)} = Wh_{t-1}^{(3)} - \frac{1}{3}(Wh_{t-1}^{(1)} + W_{\mathrm{in}}x_t)^{\odot 3} = Wh_{t-1}^{(3)} - \frac{1}{3}(h_t^{(1)})^{\odot 3}, \tag{C.29}$$

where the second equality is the first-order identity.

Uniqueness of the stationary solution of (C.25) and the convergence of (C.26) follow from the convergence convention of Appendix A. By induction, the forcing term is a polynomial of degree $2k+1$ in the linear filter $h^{(1)}$, so finite moments of order $2(2k+1)$ suffice for the L^2 statement, and the degree count coincides with the weight count. □

Lemma C.6 (Derivatives of the stationary branch on the L^p scale)**.** *Let Assumption 2 hold, let $(x_t)_{t\in\mathbb{Z}}$ be strictly stationary with $x_0 \in L^r(P)$ for every finite r, and let $(h_t^{\star,\lambda})$ be the stationary branch of Proposition C.4. Write $\|v\|_{L^p(D)} := (\mathbb{E}\|v\|_D^p)^{1/p}$. Then the following hold.*

1. (Derivatives.) *For every finite $m \ge 1$ and every $p \in [1,\infty)$ the map $\lambda \mapsto h_0^{\star,\lambda}$ is C^m from $\mathbb{R}$ into $L^p(P;\mathbb{R}^n)$. For each λ its derivatives $u_{t,j}(\lambda) := \partial_\lambda^j h_t^{\star,\lambda}$ are the unique stationary causal solutions with $\mathbb{E}\|u_{0,j}\|_D < \infty$ of*

$$u_{t,j} = A_t^\lambda W u_{t-1,j} + R_{t,j}^\lambda, \qquad A_t^\lambda := \mathrm{diag}\big(\mathrm{sech}^2 z_{t,i}^\lambda\big), \quad z_t^\lambda := Wh_{t-1}^{\star,\lambda} + \lambda W_{\mathrm{in}}x_t, \tag{C.30}$$

   where $R_{t,1}^\lambda = A_t^\lambda W_{\mathrm{in}}x_t$ and, for $j \ge 2$, $R_{t,j}^\lambda$ is a finite polynomial in x_t and $u_{t-1,1},\dots,u_{t-1,j-1}$ whose coefficients are derivatives of $\tanh$ evaluated at z_t^λ.
2. (Bounds on the L^r scale.) *The first derivative satisfies*

$$\|u_{0,1}(\lambda)\|_{L^r(D)} \le (1-q)^{-1}\|D^{1/2}W_{\mathrm{in}}\|_{\mathrm{op}}\|x_0\|_{L^r}, \tag{C.31}$$

   and for every $j \le r$ there is $C_j < \infty$, depending only on j, D, n, q and $\|D^{1/2}W_{\mathrm{in}}\|_{\mathrm{op}}\|x_0\|_{L^r}$, with

$$\sup_{\lambda\in\mathbb{R}} \|u_{0,j}(\lambda)\|_{L^{r/j}(D)} \le C_j. \tag{C.32}$$

   Each order contributes one factor $(1-q)^{-1}$.
3. (Taylor expansion.) *$h_t^{\star,-\lambda} = -h_t^{\star,\lambda}$ almost surely, so every even-order derivative vanishes at $\lambda = 0$, and $u_{t,j}(0) = j!\,h_t^{(j)}$ with $h_t^{(j)}$ the formal coefficient of Lemma C.5. Consequently, for every $K \ge 0$ and $p \in [1,\infty)$ there is a constant $C_{K,p} < \infty$, built from one factor $(1-q)^{-1}$ per derivative order and the input norm $\|x_0\|_{L^{(2K+3)p}}$, such that*

$$\sup_t \Big\| h_t^{\star,\lambda} - \sum_{k=0}^{K} \lambda^{2k+1} h_t^{(2k+1)} \Big\|_{L^p} \;\le\; C_{K,p}\,|\lambda|^{2K+3}, \qquad \lambda \in \mathbb{R}. \tag{C.33}$$

   The remainder here is the unique stationary causal one, because the branch is unique by Proposition C.4(1) and the coefficients are unique by Lemma C.5.

Only stationarity of (x_t) is used. Existence and uniqueness of the branch are given by Proposition C.4(1), whose proof is pathwise.

Idea. Steps (i) and (ii) show that every affine recursion $u_t = A_t W u_{t-1} + g_t$ with diagonal $0 \preceq A_t \preceq I$ has one stationary solution, bounded by $(1-q)^{-1}$ times its forcing. Steps (iii) and (iv) identify the first derivative by dominating the difference quotient with an input envelope. Step (v) repeats this at each higher order, with Faà di Bruno's formula giving the forcing and Hölder's inequality placing it in $L^{r/j}$. Step (vi) uses the oddness of $\tanh$ to match the derivatives at $\lambda = 0$ with the formal coefficients, and then applies Taylor's theorem.

Proof. Estimates are made path by path and then integrated. We pass between the norms $\|\cdot\|_D$ and $\|\cdot\|_2$ through (C.15), at a cost depending only on D, and we bound componentwise products through $\|v \odot w\|_2 \le \|v\|_2\|w\|_2$, which holds pathwise. Both costs are absorbed into the constants below.

(i) The linearized recursion contracts pathwise. For every diagonal A with $0 \preceq A \preceq I$ and every $v \in \mathbb{R}^n$,

$$\|AWv\|_D \;\le\; \|Wv\|_D \;\le\; q\|v\|_D, \tag{C.34}$$

by the two facts already used in Lemma C.3. The first is monotonicity of $\|\cdot\|_D$ under componentwise domination, which holds because D is diagonal, and the second is $W^\top DW \preceq q^2 D$. The bound is uniform over paths and over λ, since A_t^λ enters only through $0 \preceq A_t^\lambda \preceq I$.

(ii) Affine recursions with this linear part. Let (A_t) be a stationary sequence of diagonal matrices with $0 \preceq A_t \preceq I$, where each matrix is a measurable function of the input past, and let (g_t) be stationary and causal with $\|g_0\|_{L^p(D)} < \infty$. Then

$$u_t \;=\; A_t W u_{t-1} + g_t \tag{C.35}$$

has exactly one stationary causal solution with $\|u_0\|_{L^p(D)} < \infty$, namely the ordered series

$$u_t = \sum_{j\ge 0} \Big(\prod_{i=0}^{j-1} A_{t-i} W\Big) g_{t-j}, \tag{C.36}$$

and it obeys the pathwise envelope and the resolvent bound

$$\|u_t\|_D \;\le\; \sum_{j\ge 0} q^j \|g_{t-j}\|_D \;=:\; \Xi_t(g), \qquad \|u_0\|_{L^p(D)} \;\le\; \|\Xi_0(g)\|_{L^p} \;\le\; \frac{\|g_0\|_{L^p(D)}}{1-q}. \tag{C.37}$$

Indeed (C.34) gives

$$\|\Big(\prod_{i<j} A_{t-i} W\Big) g_{t-j}\|_D \le q^j \|g_{t-j}\|_D, \tag{C.38}$$

and Minkowski's inequality with stationarity gives

$$\|\Xi_0(g)\|_{L^p} \le \sum_{j\ge 0} q^j \|g_{-j}\|_{L^p(D)} = (1-q)^{-1} \|g_0\|_{L^p(D)}. \tag{C.39}$$

In particular, the series converges absolutely almost surely and defines a stationary causal solution. If u and u' are two such solutions, their difference obeys $\delta_t = A_t W \delta_{t-1}$ and is stationary, so

$$\|\delta_0\|_{L^p(D)} \le q^m \|\delta_{-m}\|_{L^p(D)} = q^m \|\delta_0\|_{L^p(D)} \to 0. \tag{C.40}$$

The factor $(1-q)^{-1}$ in (C.37) is the contraction margin of Assumption 2. It is the only small denominator anywhere in the construction, and each derivative order contributes one such factor.

(iii) First order, and the envelope. Taking $A_t = A_t^\lambda$ and $g_t = A_t^\lambda W_{\mathrm{in}} x_t$ in (C.35) shows that (C.30) at $j = 1$ has a unique solution. Since $\|A_t^\lambda W_{\mathrm{in}} x_t\|_D \le \|W_{\mathrm{in}} x_t\|_D$, the envelope in (C.37) becomes

$$\|u_{t,1}\|_D \;\le\; \Xi_t \;:=\; \sum_{j\ge 0} q^j \|W_{\mathrm{in}} x_{t-j}\|_D, \qquad \|\Xi_0\|_{L^r} \;\le\; \frac{\|D^{1/2} W_{\mathrm{in}}\|_{\mathrm{op}} \|x_0\|_{L^r}}{1-q}, \tag{C.41}$$

which is the first assertion of claim 2.

(iv) The candidate is the derivative. Let $v_t^\varepsilon := \varepsilon^{-1}(h_t^{\star,\lambda+\varepsilon} - h_t^{\star,\lambda})$. Applying the mean value theorem to $\tanh$ in each coordinate, we obtain

$$v_t^\varepsilon = B_t^\varepsilon \big(W v_{t-1}^\varepsilon + W_{\mathrm{in}} x_t\big), \qquad B_t^\varepsilon = \mathrm{diag}\big(\mathrm{sech}^2 \zeta_{t,i}^\varepsilon\big), \quad 0 \preceq B_t^\varepsilon \preceq I, \tag{C.42}$$

with $\zeta_{t,i}^\varepsilon$ between $z_{t,i}^\lambda$ and $z_{t,i}^{\lambda+\varepsilon}$. By (C.34), $\|v_t^\varepsilon\|_D \le q\|v_{t-1}^\varepsilon\|_D + \|W_{\mathrm{in}} x_t\|_D$, and iterating this one-step bound gives $\sup_{\varepsilon\ne 0} \|v_t^\varepsilon\|_D \le \Xi_t$ pathwise. This is the same envelope (C.41), obtained by the same iteration.

Since $\|h_t^{\star,\lambda+\varepsilon} - h_t^{\star,\lambda}\|_D \le |\varepsilon| \Xi_t \to 0$ almost surely, we have $z_t^{\lambda+\varepsilon} \to z_t^\lambda$ and hence $B_t^\varepsilon \to A_t^\lambda$ almost surely. Subtracting (C.30) at $j = 1$, we find that the gap $\delta_t := v_t^\varepsilon - u_{t,1}$ obeys

$$\delta_t = B_t^\varepsilon W \delta_{t-1} + \big(B_t^\varepsilon - A_t^\lambda\big)\big(W u_{t-1,1} + W_{\mathrm{in}} x_t\big), \tag{C.43}$$

so (C.34) gives

$$\|\delta_0\|_D \le \sum_{j\ge 0} q^j \eta_{-j}^\varepsilon, \qquad \eta_t^\varepsilon := \|(B_t^\varepsilon - A_t^\lambda)(W u_{t-1,1} + W_{\mathrm{in}} x_t)\|_D. \tag{C.44}$$

Each $\eta_t^\varepsilon \to 0$ almost surely, and since $B_t^\varepsilon - A_t^\lambda$ is diagonal with entries in $[-1,1]$ we have $\eta_t^\varepsilon \le q\Xi_{t-1} + \|W_{\mathrm{in}}x_t\|_D =: \bar\eta_t$. By Minkowski's inequality, the geometric average $\sum_{j\ge 0} q^j \bar\eta_{-j}$ lies in L^r. Dominated convergence, applied first to the series and then to the expectation, gives $\|\delta_0\|_{L^p} \to 0$ for every $p \le r$. So $u_{0,1}(\lambda)$ is the L^p-derivative of $\lambda \mapsto h_0^{\star,\lambda}$.

(v) Higher orders: Faà di Bruno and the Hölder bookkeeping. Differentiating $h_{t,i}^{\star,\lambda} = \tanh(z_{t,i}^\lambda)$ j times in λ and using Faà di Bruno's formula, we obtain

$$u_{t,j,i} \;=\; \sum_{\pi} \tanh^{(|\pi|)}\big(z_{t,i}^\lambda\big) \prod_{B\in\pi} \big(\partial_\lambda^{|B|} z_t^\lambda\big)_i, \tag{C.45}$$

where π runs over the set partitions of the j derivative slots $\{1,\dots,j\}$. Here $\partial_\lambda z_t^\lambda = W u_{t-1,1} + W_{\mathrm{in}}x_t$, and $\partial_\lambda^i z_t^\lambda = W u_{t-1,i}$ for $i \ge 2$. The one-block partition contributes $A_t^\lambda W u_{t-1,j}$, which is the linear term of (C.30). Every other partition has all blocks of size at most $j-1$, and those terms are collected into $R_{t,j}^\lambda$.

Fix r and suppose inductively that $u_{\cdot,i}$ is controlled in $L^{r/i}$ for every $i < j$. In the term indexed by π, the factor attached to a block B is a derivative of order $|B|$ and so lies in $L^{r/|B|}$. The multiplier $\tanh^{(|\pi|)}(z_{t,i}^\lambda)$ is bounded by an absolute constant, uniformly over paths and over λ, because every derivative of $\tanh$ is bounded on $\mathbb{R}$. Hölder's inequality applies to the product with the exponents $|B|/r$, $B \in \pi$, and these add up to

$$\sum_{B\in\pi} \frac{|B|}{r} \;=\; \frac{j}{r} \tag{C.46}$$

because the blocks of π partition the j slots. Hence $R_{t,j}^\lambda \in L^{r/j}$, with norm at most the product of the individual block bounds, whatever the dependence between blocks.

Since $R_{\cdot,j}^\lambda \in L^{r/j}$, step (ii) applied with $g = R_{\cdot,j}^\lambda$ produces the unique stationary causal $u_{\cdot,j}$ in that space, together with $\|u_{0,j}\|_{L^{r/j}(D)} \le (1-q)^{-1}\|R_{0,j}^\lambda\|_{L^{r/j}(D)}$. This is claim 2, where the constant C_j is built from the block constants and one further factor $(1-q)^{-1}$. The constant is uniform in λ because the only λ-dependent quantities entering it are A_t^λ and the derivatives of $\tanh$ at z_t^λ, and all of these are bounded by absolute constants.

The proof that $u_{\cdot,j}$ is the derivative of $u_{\cdot,j-1}$ is the same as in (iv). The difference quotient of $u_{\cdot,j-1}$ satisfies the same recursion up to a forcing term that tends to zero almost surely. By the induction hypothesis and the envelope (C.41), this forcing term is dominated by a fixed element of $L^{r/j}$. The same domination gives continuity of $\lambda \mapsto u_{0,j}(\lambda)$, so the map is C^m into L^p whenever $r \ge mp$, which is claim 1.

(vi) Oddness and the Taylor remainder. Since $\tanh$ is odd, $(-h_t^{\star,\lambda})$ is a bounded stationary causal solution of the recurrence at $-\lambda$ on the same input path. Uniqueness in Proposition C.4(1) therefore gives $h_t^{\star,-\lambda} = -h_t^{\star,\lambda}$ almost surely, and the even-order derivatives of the L^p-valued map $\lambda \mapsto h_0^{\star,\lambda}$ vanish at $\lambda = 0$. At this point the branch is $h_t^{\star,0} \equiv 0$, hence $A_t^0 = I$, and (C.30) divided by $j!$ is exactly the order-matching recursion (C.25). The stationary solution of that recursion is unique (Lemma C.5), so $u_{t,j}(0) = j!\, h_t^{(j)}$, and the even coefficients vanish in both descriptions.

We then apply Taylor's theorem with Bochner integral remainder to a C^{2K+3} map into L^p, together with that vanishing and the λ-uniform bound of claim 2 at order $2K+3$ (take $r = (2K+3)p$). This gives (C.33), with $C_{K,p}$ a fixed multiple of $C_{2K+3}/(2K+3)!$ that depends only on D. Stationarity makes the bound uniform in t. □

C.2.1 Proof of Theorem 3 (Graded orders at a fixed span)

The proof uses Lemmas C.5 and C.6 and the ridge-propagation Lemma D.11 of Appendix D.6.

Idea. Shared moments up to order $2k$ make the lower coefficients of M and Σ_h agree, Lemma D.11 carries the first differing coefficient to the readout, and the moment–cumulant formula reduces this coefficient at $W = 0$ to $c_k(s)$. Analyticity in (W, W_{in}) then gives genericity.

Proof. (a) By Lemma C.5 the state has the odd expansion

$$h_t = \sum_{j\le k} \lambda^{2j+1} h_t^{(2j+1)} + O(\lambda^{2k+3}), \tag{C.47}$$

where $h_t^{(2j+1)}$ is a homogeneous polynomial of degree $2j+1$ in the input path, and the remainder is controlled in every L^p by Lemma C.6. The coefficients of λ^{2j+1} in M and of λ^{2j} in Σ_h are

$$\mathbb{E}\big[x_{t+s}h_t^{(2j+1)\top}\big] \qquad \text{and} \qquad \sum_{a+b=j-1}\mathbb{E}\big[h_t^{(2a+1)}h_t^{(2b+1)\top}\big], \tag{C.48}$$

an absolutely convergent linear combination of joint moments of order $2j+2$ and a combination of joint moments of order $2j$, respectively. At any order K, Hölder's inequality turns the L^p remainder of Lemma C.6 into a remainder $O(\lambda^{2K+3})$ in M and $O(\lambda^{2K+4})$ in Σ_h, which gives (6).

The classes share every joint moment of order at most $2k$, so the coefficients of M agree up to λ^{2k-1} and those of Σ_h up to λ^{2k}. This gives the two expansions, and Lemma D.11 with $r = 2k+1$ gives $\Delta\bar{W}$.

(b) At $W = 0$ the state is $\tanh(\lambda W_{\text{in}}x_t)$, whose coefficient of λ^{2k+1} is $\tau_{2k+1}W_{\text{in}}^{\odot(2k+1)}x_t^{2k+1}$, so

$$(\Delta M^{(2k+1)})_i = \tau_{2k+1}w_i^{2k+1}\Delta\mathbb{E}[x_{t+s}x_t^{2k+1}]. \tag{C.49}$$

By the moment–cumulant formula, $\mathbb{E}[x_{t+s}x_t^{2k+1}]$ is the sum over the partitions of its $2k+2$ arguments of the products of the joint cumulants of the blocks. The partition with one block gives the joint cumulant. A partition with a block of $2k+1$ arguments has a singleton block, whose cumulant is the mean, zero. Every other partition has blocks of at most $2k$ arguments, whose cumulants are functions of joint moments of order at most $2k$ and agree between the classes. Hence $\Delta\mathbb{E}[x_{t+s}x_t^{2k+1}] = c_k(s)$.

We next show that the coefficient is real-analytic in the reservoir. Let

$$\Omega := \{(W, W_{\text{in}}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n : \rho(W) < 1\} \tag{C.50}$$

and let $\Omega^{\mathbb{C}}$ be its complex counterpart. This set is open because eigenvalues depend continuously on the entries. It is star-shaped about the origin since $\rho(sW) = |s|\rho(W)$, and hence it is connected.

By (C.26), $h_t^{(2k+1)}$ is a sum of nested Neumann series, and truncating each at depth N makes $\mathbb{E}[x_{t+s}h_t^{(2k+1)}]$ a polynomial in the entries of (W, W_{in}). On a compact $K \subset \Omega^{\mathbb{C}}$, choose $\xi < 1$ with $\rho(W) < \xi$ on K. The resolvent representation

$$W^m = \frac{1}{2\pi i}\oint_{|z|=\xi} z^m(zI - W)^{-1}dz \tag{C.51}$$

gives $\sup_K \|W^m\| \le C_K\xi^m$, and every moment in the expansion is bounded by $\mathbb{E}|x_t|^{2k+2}$ by Hölder's inequality. The discarded tails are therefore bounded by geometric series uniformly on K, and the truncations converge locally uniformly. Since each truncation is holomorphic, the limit is holomorphic on $\Omega^{\mathbb{C}}$ by Weierstrass' convergence theorem. Its coefficients are real, so it is real-analytic on Ω (Krantz & Parks, 2002). The complexification is needed because locally uniform limits of real polynomials need not be real-analytic.

The class difference of two such functions is real-analytic as well. By the computation at $W = 0$ above its i-th entry at $(0, W_{\text{in}})$ is $\tau_{2k+1}c_k(s)w_i^{2k+1}$. Here

$$\tau_{2k+1} = 2^{2k+2}(2^{2k+2} - 1)B_{2k+2}/(2k+2)!, \tag{C.52}$$

where B_{2j} are the Bernoulli numbers, which are non-zero for every $j \ge 1$ (Olver et al., 2010, 4.33.3 and 25.6.2). The entry is therefore non-zero whenever $w_i \ne 0$ and $c_k(s) \ne 0$. The zero set of a real-analytic function that is not identically zero on the connected open set Ω is Lebesgue-null (Krantz & Parks, 2002).

It remains to treat the orthogonal family, which is itself Lebesgue-null in $\mathbb{R}^{n\times n}$. Fix Q and an index i with $w_i \ne 0$. The i-th entry along the segment $\rho \mapsto (\rho Q, W_{\text{in}})$ is real-analytic in ρ on $(-1, 1)$ and non-zero at $\rho = 0$. Its zeros in $(0, 1)$ are therefore isolated, and hence countable. □

On the reversal pairs built from the m-point Gauss–Hermite innovation law, whose moments agree with the Gaussian ones up to order $2m-1$, the two classes share every joint moment of order at most $2m-2$ and differ first at order $2m$. We compute the readout difference exactly by enumerating the innovation windows, with 60 significant digits. The log-log slope of $\|\Delta\bar{W}\|$ at $\lambda \in \{0.005, 0.01, 0.02, 0.04\}$ is between 2.997 and 3.000, 4.992 and 5.000, and 6.987 and 7.000 for $m = 2, 3, 4$, both for one memoryless unit and for two units with the recurrent edge $W_{12} = 0.4$.

D SIGNAL AND NOISE AT FINITE LENGTH: PROOFS

This appendix proves Theorem 4 and the results of §4 around it. Appendix D.1 gives a dependence lemma, Appendix D.2 the cubic case at any input dimension, Appendix D.3 the size of its constant, Appendix D.4 the scale axis at finite memory and at saturation, Appendix D.5 the uniform convergence behind the lower bound, and Appendix D.6 the proof of the theorem. The proof of Theorem 4 uses Lemma D.1, Proposition D.2(a,b), Lemma D.10 and Lemma D.11.

D.1 A DEPENDENCE LEMMA

The limit laws of this appendix and of Appendix E use one common fact. The reservoir statistics are bounded functionals of the innovations, and their dependence on any single innovation decays geometrically. The following lemma states this fact once, in the form that the central limit theorem of Wu (2005) requires.

Lemma D.1 (Summable dependence of the reservoir statistics)**.** *Let*

$$x_t = \sum_{j=0}^{m} \vartheta_j e_{t-j} \tag{D.1}$$

with $\vartheta_j \in \mathbb{R}^{d_x \times d_e}$ and iid $\|e_t\|_2 \le B_e$, so that

$$\|x_t\|_2 \le B_x := B_e \sum_j \|\vartheta_j\|_{\mathrm{op}}, \tag{D.2}$$

let Assumption 2 hold, and fix a span s.

For a stationary sequence Y of functionals of the innovations, let $\delta_2(j) := \|Y_j - Y_j^*\|_{L^2}$ be its functional dependence measure, where Y^* denotes the sequence obtained when the innovation e_0 is replaced by an independent copy.

(i) *At fixed λ, with $h^\star$ the stationary* tanh *state, the entries of $Y_t := (x_{t+s} h_t^{\star\top},\, h_t^\star h_t^{\star\top})$ satisfy $\delta_2(j) \le C q^j$.*

(ii) *With $a^\star$ and $c^\star$ the stationary linear and cubic states, the entries of*

$$Z_t := (x_{t+s} a_t^{\star\top},\, x_{t+s} c_t^{\star\top},\, a_t^\star a_t^{\star\top}) \tag{D.3}$$

satisfy $\delta_2(j) \le C(j+1) q^{j/2}$.

(iii) *In both cases the dependence is summable. Consequently, for every fixed linear combination of the entries, and jointly by the Cramér–Wold device, $\sqrt{L}\,(\bar{Y}_L - \mathbb{E}Y_0) \Rightarrow \mathcal{N}(0, \Omega)$, where Ω is the long-run covariance. Moreover $L\,\mathrm{Cov}(\bar{Y}_L) \to \Omega$ and $\mathbb{E}\bar{Y}_L = \mathbb{E}Y_0 + O(1/L)$. Along the trajectories from $h_0 = 0$, the window averages of the same statistics differ pathwise from those of the stationary ones by $O(1/L)$, and the $\sqrt{L}$ scaling removes this difference.*

Proof. Idea. Replacing one innovation changes the input at finitely many times, and the contraction of Assumption 2 makes the effect on the state decay geometrically. This gives (i). The linear and cubic states of (ii) are sums over all past inputs, so they are truncated at a lag J. The truncated statistic does not depend on e_0 at large lags, and choosing J to balance the truncation error introduces the factor $j+1$ and the square root in the rate. Part (iii) is the central limit theorem of Wu (2005).

(i) Consider two trajectories driven by input paths that differ only at times in a set T_0. Iterating Lemma C.3 with the drive difference gives

$$\|h_t - h'_t\|_D \le \sum_{\tau \in T_0, \tau \le t} q^{t-\tau} |\lambda| \|W_{\mathrm{in}}\|_D \|x_\tau - x'_\tau\|_2, \tag{D.4}$$

where $\|W_{\mathrm{in}}\|_D$ is the operator norm from $(\mathbb{R}^{d_x}, \|\cdot\|_2)$ to $(\mathbb{R}^n, \|\cdot\|_D)$. Replacing e_0 changes x at each of the times $0, \dots, m$ by at most $2B_e\|\vartheta_j\|_{\mathrm{op}}$, so $\|h_t - h'_t\|_D \le C q^{t-m}$ for $t > m$. The entries of Y_t are bounded, since $|h| \le 1$ and $\|x\|_2 \le B_x$, and they are Lipschitz in h. This gives the bound, with a constant that absorbs the finitely many early terms.

(ii) Truncate the Neumann series of $a^\star$ and of $c^\star$ at lag J. Every entry of the truncated $Z_t^{(J)}$ is a finite sum of products of at most four inputs and depends only on the innovations at lags at most $2J + m + s$. The truncation error is uniform in t: by (C.34),

$$\|a_t^\star - a_t^{\star,(J)}\|_D \le q^J \|W_{\mathrm{in}}\|_D B_x/(1-q), \tag{D.5}$$

and cubing is Lipschitz on the bounded range of $a^\star$. The discarded tail of the cubic series and the perturbed cubes inside the retained terms therefore together give $\|c_t^\star - c_t^{\star,(J)}\|_D \le C(J+1)q^J$. Since norms are equivalent in finite dimension, $\|Z_t - Z_t^{(J)}\|_{L^2} \le C'(J+1)q^J$.

Replacing e_0 leaves $Z_j^{(J)}$ unchanged as soon as $j > 2J + m + s$, so $\delta_2(j) \le 2C'(J+1)q^J$ for every such J. Taking $J = \lfloor (j-m-s-1)/2 \rfloor$ gives $\delta_2(j) \le C''(j+1)q^{j/2}$.

(iii) In both cases $\sum_j \delta_2(j) < \infty$, so the central limit theorem of Wu (2005) applies to each fixed linear combination of the entries. The Cramér–Wold device then gives the joint limit, whose covariance is the sum of the autocovariances. Boundedness together with summable dependence gives $L\,\mathrm{Cov}(\bar{Y}_L) \to \Omega$ and $\mathbb{E}\bar{Y}_L = \mathbb{E}Y_0 + O(1/L)$, where the bias comes from the window edges.

The trajectories from $h_0 = 0$ differ from the stationary ones by $O(q^t)$ for h and a (Proposition C.4(2) and (C.34)). For c they differ by $O((t+1)q^t)$, because its recursion is driven by a term of that size. Both bounds are summable, so the window averages differ by $O(1/L)$.

The two limits in (ii) are not interchanged. The quantity $C'(J+1)q^J$ bounds an unnormalized average and is not $o(L^{-1/2})$ at fixed J. For this reason the limit is obtained from the dependence measure rather than from a finite-dependence argument. □

Write $h_t^{(1)}$ and $h_t^{(3)}$ for the terms of order λ and λ^3 of the state (Lemma C.5), $\hat{M}_1$ and $\hat{M}_{3,0}$ for their sample cross-moments with the target, $\hat{\Sigma}_{11}$ for the sample covariance of $h_t^{(1)}$ and Σ_{11} for its population value, and set

$$\hat{R}_\lambda := (\gamma I + \lambda^2 \hat{\Sigma}_{11})^{-1}, \qquad \bar{V}_1 := \tfrac{1}{2}\big(\lim_L L\,\mathrm{Cov}_1 \hat{M}_1 + \lim_L L\,\mathrm{Cov}_2 \hat{M}_1\big), \tag{D.6}$$

with R_λ the population form of $\hat{R}_\lambda$ and $\bar{V}_1$ the pooled long-run covariance of the linear term.

Proposition D.2 (Signal and noise at finite L)**.** *Fix $\gamma > 0$ and a span s, and let Assumption 2 hold. Let $x_t = \sum_{j=0}^m \vartheta_j e_{t-j}$ with iid, centred and bounded innovations $e_t \in \mathbb{R}^{d_e}$ and coefficients $\vartheta_j \in \mathbb{R}^{d_x \times d_e}$. Matrices are identified with their vectorizations.*

(a) Expansion. *Pathwise, uniformly in L, and for λ below a threshold set by the contraction constants,*

$$\hat{W} = \lambda\, \hat{M}_1 \hat{R}_\lambda + \lambda^3\, \hat{M}_{3,0} \hat{R}_\lambda + O(\lambda^5), \qquad \mathbb{E}\hat{M}_{3,0} = M^{(3)}_{g_s,P} + O(1/L). \tag{D.7}$$

(b) Noise. *The limiting covariance of the readout is*

$$\lim_{L\to\infty} L\,\mathrm{Cov}(\hat{W}) = \lambda^2 R_\lambda V_1 R_\lambda + O(\lambda^4), \tag{D.8}$$

where V_1 is the long-run covariance of the sample autocovariances carried through the Krylov weights $W^k W_{\mathrm{in}}$, given by Bartlett's formula (D.18) *at $d_x = 1$. The $O(\lambda^4)$ term depends on γ.*

(c) Separation. *Let two classes share the autocovariance, with $\Delta M^{(3)} \neq 0$ and $\bar{V}_1 \succ 0$. The separation of their mean readouts in the metric of the pooled limiting covariance is*

$$d^2 \;:=\; L\,\Delta\mathbb{E}\hat{W}\,\big(\lim_L L\,\mathrm{Cov}\hat{W}\big)^{-1}\Delta\mathbb{E}\hat{W}^\top \;=\; \lambda^4\, L\,\Delta M^{(3)}\,\bar{V}_1^{-1}\,(\Delta M^{(3)})^\top\,(1+o(1)), \tag{D.9}$$

where $o(1)$ refers to $L \to \infty$ at fixed λ, followed by $\lambda \to 0$.

D.2 Proof of Proposition D.2 (Signal and noise at finite L)

Proof. Idea. The contraction keeps the state of order λ along every path, so $\tanh$ can be replaced by its cubic Taylor polynomial with a remainder of order λ^5, uniformly in L. This splits the moments and the readout into a linear part, a cubic part and a remainder (a). The linear part is a linear image of the sample autocovariances, and their central limit theorem gives the noise V_1 at order λ (b). Two classes with one autocovariance share the linear mean, so their mean difference starts at order λ^3, and the ratio of the two gives $\lambda^4 L$ (c).

(a) Pathwise expansion. Let $\|e_t\|_2 \le B_e$, so that $\|x_t\|_2 \le B_x := B_e \sum_j \|\vartheta_j\|_{\mathrm{op}}$. Take $\|W_{\mathrm{in}}\|_D$ as in Lemma D.1, and write $z_t = W h_{t-1} + \lambda W_{\mathrm{in}} x_t$ for the pre-activation. Since $|\tanh z| \le |z|$ componentwise and $\|\cdot\|_D$ is monotone under componentwise domination, Assumption 2 gives

$$\|h_t\|_D \le q\|h_{t-1}\|_D + \lambda\|W_{\mathrm{in}}\|_D B_x, \quad \text{hence} \quad \sup_t \|h_t\|_D \le \lambda C_0, \quad C_0 := \frac{\|W_{\mathrm{in}}\|_D B_x}{1-q}, \tag{D.10}$$

uniformly in L. Define, from zero initial states,

$$h_t^{(1)} = W h_{t-1}^{(1)} + W_{\text{in}} x_t, \qquad h_t^{(3)} = W h_{t-1}^{(3)} - \frac{1}{3}(h_t^{(1)})^{\odot 3}. \tag{D.11}$$

By the same contraction, both are bounded in $\|\cdot\|_D$ uniformly in t. The remainder $r_t := h_t - \lambda h_t^{(1)} - \lambda^3 h_t^{(3)}$ has pre-activation

$$z_t = \lambda h_t^{(1)} + \lambda^3 W h_{t-1}^{(3)} + W r_{t-1}. \tag{D.12}$$

For λ below a threshold set by C_0, the bound $|\tanh z - z + z^3/3| \le |z|^5$ for $|z| \le 1$ gives

$$\|r_t\|_D \le q(1 + C_2\lambda^2)\|r_{t-1}\|_D + C_1\lambda^5, \qquad \text{hence} \qquad \sup_t \|r_t\|_D \le C_3\lambda^5. \tag{D.13}$$

Substituting into the moments of (2),

$$\hat{\Sigma}_h = \lambda^2 \hat{\Sigma}_{11} + \lambda^4 \hat{\Sigma}_{13}^{\text{s}} + O(\lambda^6), \qquad \hat{M} = \lambda \hat{M}_1 + \lambda^3 \hat{M}_{3,0} + O(\lambda^5), \tag{D.14}$$

where $\hat{\Sigma}_{13}^{\text{s}}$ is the symmetrized sample cross-covariance of $h^{(1)}$ and $h^{(3)}$ and $\hat{M}_{3,0}$ is the sample cross-moment of x_{t+s} with $h_t^{(3)}$. Every constant is uniform in L.

Since $\|\hat{R}_\lambda\|_{\text{op}} \le \gamma^{-1}$, the resolvent identity gives

$$(\hat{\Sigma}_h + \gamma I)^{-1} = \hat{R}_\lambda - \lambda^4 \hat{R}_\lambda \hat{\Sigma}_{13}^{\text{s}} \hat{R}_\lambda + O(\lambda^6), \tag{D.15}$$

and multiplying out yields $\hat{W} = \lambda \hat{M}_1 \hat{R}_\lambda + \lambda^3 \hat{M}_{3,0} \hat{R}_\lambda + O(\lambda^5)$.

Finally, $\mathbb{E}\hat{M}_{3,0}$ equals the stationary expectation $\mathbb{E}[x_{t+s}(h_t^{(3)})^\top]$ up to the $O(1/L)$ transient from $h_0 = 0$. We write $M_{g_s,P}^{(3)}$ for that expectation, with $h^{(3)}$ the cubic coefficient of Lemma C.5, which holds at every d_x. At $d_x = 1$ it is the coefficient of Theorem 3(a), because the population expansion is the expectation of the pathwise one.

(b) The linear noise. Up to $O(1/L)$ edge terms,

$$\hat{M}_1 = \sum_{k \ge 0} \hat{\Gamma}(s+k)\, W_{\text{in}}^\top (W^\top)^k, \qquad \hat{\Gamma}(\tau) = \tfrac{1}{L} \sum_t x_{t+\tau} x_t^\top, \tag{D.16}$$

which is a linear combination of sample autocovariances with geometrically decaying weights. In vectorized form,

$$\operatorname{vec} \hat{M}_1 = \sum_{k \ge 0} (W^k W_{\text{in}} \otimes I_{d_x}) \operatorname{vec} \hat{\Gamma}(s+k). \tag{D.17}$$

By Lemma D.1 and the Cramér–Wold device, the vectorized sample autocovariances at any finite set of lags are jointly asymptotically normal.

Their covariance blocks $c_{p,p'} \in \mathbb{R}^{d_x^2 \times d_x^2}$ are the long-run cross-covariances of $\operatorname{vec}(x_{t+p} x_t^\top)$ and $\operatorname{vec}(x_{t+p'} x_t^\top)$. This proof uses only the existence of these blocks, and Lemma D.3 uses their iid form. At $d_x = 1$, for a linear process with iid innovations and $\mathbb{E}e_t^4 = \eta \sigma_e^4$, they are

$$c_{p,p'} \;=\; (\eta - 3)\Gamma(p)\Gamma(p') + \sum_{m \in \mathbb{Z}} \big[\Gamma(m)\Gamma(m - p + p') + \Gamma(m + p')\Gamma(m - p)\big] \tag{D.18}$$

(Brockwell & Davis, 1991). The geometric weights make the truncation of the sum over k uniform, so

$$\begin{aligned} &\sqrt{L} \operatorname{vec}(\hat{M}_1 - M_1) \Rightarrow \mathcal{N}(0, V_1), \\ &V_1 = \sum_{k,k'} (W^k W_{\text{in}} \otimes I_{d_x})\, c_{s+k,s+k'}\, (W^{k'} W_{\text{in}} \otimes I_{d_x})^\top. \end{aligned} \tag{D.19}$$

At $d_x = 1$ and for iid input, $c_{0,0} = \mu_4 - \sigma^4$, $c_{p,p} = \sigma^4$ for $p \ge 1$ and $c_{p,p'} = 0$ otherwise. This gives the form of V_1 displayed in Appendix F.5 at $s = 0$.

For the readout,

$$\hat{R}_\lambda - R_\lambda = -\lambda^2 R_\lambda (\hat{\Sigma}_{11} - \Sigma_{11}) \hat{R}_\lambda = O(\lambda^2/\sqrt{L}), \tag{D.20}$$

so

$$\lambda \hat{M}_1 \hat{R}_\lambda = \lambda M_1 R_\lambda + \lambda(\hat{M}_1 - M_1)R_\lambda + O(\lambda^3/\sqrt{L}),$$
$$L \operatorname{Cov}(\operatorname{vec} \hat{W}) = \lambda^2 (R_\lambda \otimes I_{d_x}) V_1 (R_\lambda \otimes I_{d_x}) + O(\lambda^4). \tag{D.21}$$

The $O(\lambda^4)$ term collects the covariance fluctuation $\lambda^3 M_1 R_\lambda(\hat{\Sigma}_{11} - \Sigma_{11})R_\lambda$ and the cubic term. Its constant depends on γ, so the displayed term is the leading noise only while $\lambda^2 \|\Sigma_{11}\|_{\mathrm{op}} \ll \gamma$. Likewise,

$$\mathbb{E}\hat{W} = \lambda M_1 R_\lambda + \lambda^3 M^{(3)} R_\lambda + O(\lambda/L) + O(\lambda^5), \tag{D.22}$$

where the $O(\lambda/L)$ term is $\lambda(\mathbb{E}\hat{M}_1 - M_1)R_\lambda$. It comes from the transient of $h^{(1)}$ from $h_0 = 0$ and from the window edges, and it is a function of the autocovariance alone. The term $\mathbb{E}[(\hat{M}_1 - M_1)(\hat{R}_\lambda - R_\lambda)]$ is $O(\lambda^3/L)$.

(c) Whitened separation. Two classes with one autocovariance share $M_1, \Sigma_{11}, R_\lambda$ and the $O(\lambda/L)$ initialization term, so by (b)

$$\Delta\mathbb{E}\hat{W} = \lambda^3 \Delta M^{(3)} R_\lambda + O(\lambda^5) + O(\lambda^3/L),$$
$$\text{pooled } L \operatorname{Cov}(\operatorname{vec} \hat{W}) = \lambda^2 \mathcal{R}_\lambda \bar{V}_1 \mathcal{R}_\lambda + O(\lambda^4), \qquad \mathcal{R}_\lambda := R_\lambda \otimes I_{d_x}. \tag{D.23}$$

Hence, with $\delta := \operatorname{vec} \Delta M^{(3)}$,

$$d^2 = \lambda^4 L \, \delta^\top \mathcal{R}_\lambda \big(\mathcal{R}_\lambda \bar{V}_1 \mathcal{R}_\lambda\big)^{-1} \mathcal{R}_\lambda \, \delta \, (1 + o(1)), \tag{D.24}$$

and the resolvents cancel because the Mahalanobis distance is invariant under invertible affine maps applied to both classes, and $\mathcal{R}_\lambda$ is symmetric and invertible. This is (D.9).

Lemma D.3 gives a controllability condition for $\bar{V}_1 \succ 0$ under iid input.

□

Lemma D.3 (Positivity of the linear noise)**.** *Let x_t be iid, zero-mean, with $\mathbb{E}\|x_t\|^4 < \infty$ and $\Sigma_x := \mathbb{E}x_t x_t^\top \succ 0$, and let $\rho(W) < 1$. Then for every span $s \geq 0$*

$$V_1 \;\succeq\; \Sigma_W \otimes \Sigma_x, \qquad \Sigma_W := \sum_{k \geq 1} W^k W_{\mathrm{in}} \Sigma_x W_{\mathrm{in}}^\top (W^\top)^k, \tag{D.25}$$

so $V_1 \succ 0$ whenever (W, WW_{in}) is controllable. The same bound holds for the pooled $\bar{V}_1$ of two classes that share Σ_x. At $s \geq 1$ the sum may start at $k = 0$, and then (W, W_{in}) controllable suffices.

Proof. Write $\operatorname{vec}(x_{t+p} x_t^\top) = x_t \otimes x_{t+p}$. For iid zero-mean input the sequences $(x_t \otimes x_{t+p})_t$, $p \geq 0$, are uncorrelated across times and across $p \neq p'$. Indeed, a moment $\mathbb{E}[(x_t \otimes x_{t+p})(x_{t'} \otimes x_{t'+p'})^\top]$ of four zero-mean inputs vanishes unless every time appears at least twice, and when $p \geq 1$ or $p' \geq 1$ this forces $t = t'$ and $p = p'$. The blocks of (b) are therefore $c_{p,p'} = 0$ for $p \neq p'$, $c_{p,p} = \Sigma_x \otimes \Sigma_x$ for $p \geq 1$, and $c_{0,0} = \operatorname{Cov}(x_t \otimes x_t) \succeq 0$.

Consequently

$$V_1 = \sum_{k \geq 0} (W^k W_{\mathrm{in}} \otimes I) \, c_{s+k,s+k} \, (W^k W_{\mathrm{in}} \otimes I)^\top$$
$$\succeq \sum_{k \geq 1} (W^k W_{\mathrm{in}} \otimes I)(\Sigma_x \otimes \Sigma_x)(W^k W_{\mathrm{in}} \otimes I)^\top = \Sigma_W \otimes \Sigma_x, \tag{D.26}$$

by $(A \otimes I)(\Sigma \otimes \Sigma)(A \otimes I)^\top = (A\Sigma A^\top) \otimes \Sigma$. At $s \geq 1$ the term $k = 0$ has the same form and can be kept. The sum converges since $\rho(W) < 1$.

The Kronecker product of two positive definite matrices is positive definite. Given $\Sigma_x \succ 0$, we have $\Sigma_W \succ 0$ exactly when the vectors $W^k W_{\mathrm{in}} c$ with $k \geq 1$ and $c \in \mathbb{R}^{d_x}$ span $\mathbb{R}^n$, that is, when (W, WW_{in}) is controllable. Pooling two classes with the same Σ_x averages their $c_{0,0}$ blocks and leaves the bound unchanged.

At $d_x > 1$ the lag-zero block cannot replace the condition. Since $x_t \otimes x_t$ is the vectorization of a symmetric matrix, $c_{0,0}$ annihilates the vectorization of every antisymmetric matrix and has rank at most $d_x(d_x + 1)/2 < d_x^2$. In particular, at $W = 0$ the matrix $V_1 = (W_{\mathrm{in}} \otimes I) c_{s,s} (W_{\mathrm{in}} \otimes I)^\top$ has rank at most d_x^2, which is below $d_x n$ once $n > d_x$. A memoryless reservoir therefore has a singular V_1 at $d_x > 1$. □

Remark D.4 (The relative ridge). *Under the relative ridge $\gamma = g\,\mathrm{tr}\hat{\Sigma}_h/n$ the penalty scales with λ^2 as the state covariance does, and part (a) becomes*

$$\lambda\hat{W} = F(\hat{M}_1, \hat{\Sigma}_{11}) + O(\lambda^2), \qquad F(M,S) := M\big(S + g\,\mathrm{tr}(S)\,I/n\big)^{-1}. \tag{D.27}$$

For two classes that share the autocovariance write $\Sigma_{h,y} = \lambda^2 S + \lambda^4 U_y + O(\lambda^6)$ and $M_y = \lambda M_1 + \lambda^3 M_y^{(3)} + O(\lambda^5)$, with $S = \Sigma_{11}$ and M_1 common, and set $S_g := S + g\,\mathrm{tr}(S)I/n$ and $U_{g,y} := U_y + g\,\mathrm{tr}(U_y)I/n$. Then

$$\lambda\bar{W}_y = M_1 S_g^{-1} + \lambda^2\big(M_y^{(3)} S_g^{-1} - M_1 S_g^{-1} U_{g,y} S_g^{-1}\big) + O(\lambda^4), \tag{D.28}$$

so the class signal at order λ^2 is $\delta_{\mathrm{rel}} := \mathrm{vec}\big(\Delta M^{(3)} S_g^{-1} - M_1 S_g^{-1}(\Delta U_g) S_g^{-1}\big)$, which contains the cubic correction of the state covariance. Let V_{rel}/L be the leading pooled covariance of $\lambda\hat{W}$, the delta-method covariance of F at the population autocovariances, so that the fluctuations are of order $1/\sqrt{L}$. If $\delta_{\mathrm{rel}} \neq 0$ and $V_{\mathrm{rel}} \succ 0$, the squared separation is $\lambda^4 L\,\delta_{\mathrm{rel}}^\top V_{\mathrm{rel}}^{-1}\delta_{\mathrm{rel}}(1+o(1))$ in the order of limits of Proposition D.2(c), by the argument of part (c) with the delta method applied to F. Neither condition is automatic, and this is a statement about the separation, not a lower bound for the relative ridge. For example, with $n = 1$, $W = 0$, $W_{\mathrm{in}} = 1$ and the self-step target, $F \equiv 1/(1+g)$ and the leading noise vanishes. For two iid laws with one variance, different fourth moments and non-constant x_t^2, the squared separation then grows as L at small scale. Appendix F.5 measures the covariance on one setting.

D.3 The Krylov structure of the constant

The constant of the separation law is finite when $\bar{V}_1 \succ 0$, but $\bar{V}_1$ can have very small eigenvalues. The linear term is a linear image of the sample autocovariances through the Krylov vectors $W^k W_{\mathrm{in}}$, whose norms decay geometrically, and this bounds the linear noise outside the span of the first few of them.

Proposition D.5 (The linear noise outside the leading Krylov directions). *Let $d_x = 1$, fix a span s, and let $\|W^k\|_2 \le Cq^k$ for all $k \ge 0$, with $q < 1$. Write*

$$\hat{M}_1 = \sum_{k\ge 0} \hat{\gamma}_{s+k}\, W_{\mathrm{in}}^\top (W^\top)^k \tag{D.29}$$

with $\hat{\gamma}_j$ the sample autocovariance at lag j, and let Γ^{B} be the pooled limiting covariance of $\sqrt{L}(\hat{\gamma}_{s+k})_{k\ge 0}$, a bounded operator on ℓ^2 for the moving averages of Proposition D.2. Let $\mathcal{K}_\ell := \mathrm{span}\{W^k W_{\mathrm{in}} : k < \ell\}$ and let Π_ℓ be the orthogonal projection onto $\mathcal{K}_\ell^\perp$.

(a) *For every unit vector $u \perp \mathcal{K}_\ell$,*

$$u^\top \bar{V}_1 u \le \varepsilon_\ell := \|\Gamma^{\mathrm{B}}\|\, C^2 \|W_{\mathrm{in}}\|^2 q^{2\ell}/(1-q^2). \tag{D.30}$$

Hence $\bar{V}_1$ has at most ℓ eigenvalues above ε_ℓ.

(b) *If $\bar{V}_1 \succ 0$, the constant $\kappa_\star := \Delta M^{(3)} \bar{V}_1^{-1} \Delta M^{(3)\top}$ of Proposition D.2(c), which is κ_1 of Theorem 4 when the classes share V_1 at $d_x = 1$, satisfies*

$$\kappa_\star \ge \|\Pi_\ell \Delta M^{(3)\top}\|^2/\varepsilon_\ell \qquad \text{for every } \ell < n. \tag{D.31}$$

Proof. (a) Let $K : \ell^2 \to \mathbb{R}^n$ map the k-th unit vector to $W^k W_{\mathrm{in}}$, so that $\bar{V}_1 = K\Gamma^{\mathrm{B}} K^\top$. For $u \perp \mathcal{K}_\ell$ the coordinates $(K^\top u)_k = u^\top W^k W_{\mathrm{in}}$ vanish for $k < \ell$ and are at most $Cq^k\|W_{\mathrm{in}}\|$ in absolute value for $k \ge \ell$. Hence

$$u^\top \bar{V}_1 u \le \|\Gamma^{\mathrm{B}}\|\|K^\top u\|^2 \le \|\Gamma^{\mathrm{B}}\| C^2 \|W_{\mathrm{in}}\|^2 \sum_{k\ge\ell} q^{2k} = \varepsilon_\ell. \tag{D.32}$$

The subspace $\mathcal{K}_\ell^\perp$ has dimension at least $n - \ell$, so by the Courant–Fischer theorem the $(\ell+1)$-th largest eigenvalue of $\bar{V}_1$ is at most ε_ℓ.

(b) Write $\delta = \Delta M^{(3)\top}$. By the Cauchy–Schwarz inequality in the inner product of $\bar{V}_1^{-1}$, $(u^\top\delta)^2 \le (u^\top \bar{V}_1 u)\,\kappa_\star$ for every u. Take $u = \Pi_\ell \delta$. Then $u^\top\delta = \|\Pi_\ell\delta\|^2$ and, by (a), $u^\top\bar{V}_1 u \le \varepsilon_\ell\|\Pi_\ell\delta\|^2$. □

Part (b) gives a lower bound on the constant. When the cubic signal, an entrywise power of Krylov vectors, has a non-zero component outside $\mathcal{K}_{n-1}$, as on all the reservoirs of Appendix F.6, $\kappa_\star$ is at least of order $q^{-2(n-1)}$ times the squared norm of that component. The bound of part (a) is not tight. On the reservoirs of Appendix F.6 the eigenvalues of $\bar{V}_1$ lie one to two orders of magnitude below ε_ℓ, and for $\rho \le 0.9$ and large n the computed $\kappa_\star$ grows roughly at the rate ρ^{-2n}.

D.4 THE SCALE AXIS: FINITE MEMORY AND SATURATION

Proposition D.6 (The scale axis with finite memory)**.** *Let the graph of W have no path longer than p, that is, $|W|^{p+1} = 0$ entrywise, with $W = 0$ the case $p = 0$. Let the washout be at least p, and write $\varpi_t := (x_t, \dots, x_{t-p})$ for the window of the last $p + 1$ inputs.*

(a) Window dependence. *At every scale the state is $h_t = \mathfrak{h}_\lambda(\varpi_t)$ for a fixed map $\mathfrak{h}_\lambda$. The population readout therefore depends on P only through the law of (x_{t+s}, ϖ_t), and two processes with one such law share it at every scale and every ridge.*

(b) A fixed regression. *Let x_t take finitely many values, and define*

- *the frame F_λ, with one column $\mathfrak{h}_\lambda(\varpi)$ for each value ϖ of the window, and its Gram matrix $G_\lambda := F_\lambda^\top F_\lambda$,*
- *the diagonal matrix $\hat{\Lambda}$ of the window frequencies over the fit window,*
- *the matrix $\hat{Y}$ whose column ϖ is $\frac{1}{L}\sum_{t \in I_s} x_{t+s}\mathbf{1}\{\varpi_t = \varpi\}$.*

Then, exactly,

$$\hat{W} = \hat{Y}\,(G_\lambda \hat{\Lambda} + \gamma I)^{-1} F_\lambda^\top = \hat{\beta}\, G_\lambda^+ F_\lambda^\top, \tag{D.33}$$

where the window statistic $(\hat{\Lambda}, \hat{Y})$ depends on neither λ, W, W_{in} nor γ, and the coefficient $\hat{\beta}$ is the ridge regression of x_{t+s} on the window indicators with penalty $\gamma\,\mathrm{tr}(\beta G_\lambda^+ \beta^\top)$, over coefficients β whose rows lie in the range of G_λ.

(c) Exact computation. *Let x_t moreover be a moving average of finite order of iid innovations with finitely many values. The population readout and the limiting covariance of $\sqrt{L}\,\hat{W}$ are then finite sums over windows of innovations, so the separation per unit length, taken in the pseudo-inverse of the pooled covariance, and the one-step R^2 are computable exactly at every scale.*

D.4.1 PROOF OF PROPOSITION D.6

Proof. Idea. Part (a) is an induction on the longest path into each unit, which shows that every unit depends on a bounded window of inputs. The state is then the column of a fixed frame F_λ selected by the window, so every moment factors through the window frequencies, and part (b) is linear algebra on this factorization. Part (c) follows because every quantity is a finite sum over innovation windows.

(a) Let $d(i)$ be the length of the longest path into unit i in the graph with an edge $j \to i$ whenever $W_{ij} \neq 0$. An entry of $|W|^k$ is non-zero exactly when a path of length k joins the two units, so $|W|^{p+1} = 0$ gives $d(i) \leq p$. We show by induction on $d(i)$ that $h_{t,i}$ is a function of $(x_t, \dots, x_{t-d(i)})$ for every $t \geq d(i) + 1$, whatever λ. If $d(i) = 0$, the unit has no incoming edge and $h_{t,i} = \tanh(\lambda(W_{\text{in}} x_t)_i)$. If $d(i) = k \geq 1$, every j with $W_{ij} \neq 0$ has $d(j) \leq k - 1$, so $h_{t-1,j}$ is a function of $(x_{t-1}, \dots, x_{t-k})$ for $t - 1 \geq d(j) + 1$, and

$$h_{t,i} = \tanh\Big(\sum_j W_{ij} h_{t-1,j} + \lambda(W_{\text{in}} x_t)_i\Big) \tag{D.34}$$

is a function of $(x_t, \dots, x_{t-k})$. The same induction applies to the stationary solution, which is therefore the same function of the window. With a washout of at least p, every h_t in the fit window is $\mathfrak{h}_\lambda(\varpi_t)$.

The moments $\hat{\Sigma}_h$ and $\hat{M}$ are then window averages of functions of (x_{t+s}, ϖ_t), and their population values, together with $\bar{W} = M(\Sigma_h + \gamma I)^{-1}$, are expectations under the law of this pair.

(b) Write $\mathbf{1}_\varpi$ for the indicator vector of the window value ϖ. Part (a) gives $h_t = F_\lambda \mathbf{1}_{\varpi_t}$, hence $\hat{\Sigma}_h = F_\lambda \hat{\Lambda} F_\lambda^\top$ and $\hat{M} = \hat{Y} F_\lambda^\top$, and neither $\hat{\Lambda}$ nor $\hat{Y}$ involves the reservoir or the ridge. The identity

$$F^\top (F \hat{\Lambda} F^\top + \gamma I) = (G\hat{\Lambda} + \gamma I) F^\top \tag{D.35}$$

gives the first form of (D.33). The matrix $G\hat{\Lambda} + \gamma I$ is invertible because the eigenvalues of $G\hat{\Lambda}$ are those of $\hat{\Lambda}^{1/2} G \hat{\Lambda}^{1/2} \succeq 0$.

For the second form, write the fitted value of a readout W at time t as $WF\mathbf{1}_{\varpi_t} = \beta \mathbf{1}_{\varpi_t}$ with $\beta := WF$. The coefficients β that arise in this way are those whose rows lie in the range of $F^\top$, which is the range of G. For a given such β, the readout of least Frobenius norm with $WF = \beta$ is $\beta G^+ F^\top$, and its squared norm is $\mathrm{tr}(\beta G^+ \beta^\top)$. The ridge readout minimizes the squared error

plus γ times the squared norm, so it has least norm among the readouts with its own fitted values. Hence $\hat{W} = \hat{\beta} G^{+} F^{\top}$ with $\hat{\beta} = \hat{W} F$, and $\hat{\beta}$ minimizes the squared error plus $\gamma \operatorname{tr}(\beta G^{+} \beta^{\top})$ over the admissible β.

(c) Let

$$x_t = \sum_{j=0}^{m} \vartheta_j e_{t-j} \tag{D.36}$$

with innovations taking finitely many values. The pair (x_{t+s}, ϖ_t) is a function of the innovation window $(e_{t+s}, \dots, e_{t-p-m})$, so $(\hat{\Lambda}, \hat{Y})$ is the window mean of a bounded function of a finitely dependent sequence. Its mean is a finite sum over innovation windows. Its long-run covariance is a finite sum over lags smaller than the window length, each again a finite sum over innovation windows. The readout (D.33) is a smooth function of $(\hat{\Lambda}, \hat{Y})$, so by the delta method the limiting covariance of $\sqrt{L}\,\hat{W}$ is a finite sum as well. □

Remark D.7 (Truncating the memory)**.** *Let Assumption 2 hold for an arbitrary W, and let $h_t^{[p]}$ be the state of the same reservoir restarted from zero at time $t-p-1$, so that it depends on the last $p+1$ inputs only. By Lemma C.3, applied $p+1$ times,*

$$\|h_t - h_t^{[p]}\|_D \le q^{p+1} (\operatorname{tr} D)^{1/2}, \qquad \text{so} \qquad \|h_t - h_t^{[p]}\|_2 \le \varepsilon_p := q^{p+1} (\operatorname{tr} D / \min_i D_{ii})^{1/2}. \tag{D.37}$$

Since $\|h_t\|_2 \le \sqrt{n}$, the population moments of the two reservoirs differ by at most $\mathbb{E}\|x_0\| \varepsilon_p$ in M and $2\sqrt{n}\,\varepsilon_p$ in Σ_h. The readout map is Lipschitz with constants γ^{-1} in M and $\gamma^{-2}\|M\|$ in Σ_h, and $\|M\| \le \sqrt{n}\,\mathbb{E}\|x_0\|$. The population readouts therefore differ by at most

$$\mathbb{E}\|x_0\|\, \varepsilon_p (\gamma^{-1} + 2n\gamma^{-2}), \tag{D.38}$$

uniformly in λ. Two processes whose pairs (x_{t+s}, ϖ_t) have one law thus have population readouts within twice this bound of each other at every scale. The bound concerns part (a) only. The exact form (D.33) *and the covariance of part (c) do not transfer, since an error of order q^{p+1} is not small against fluctuations of order $L^{-1/2}$.*

Proposition D.8 (The saturation limit)**.** *Let $(W_{\mathrm{in}} x_t)_i \neq 0$ almost surely for every i. For any fixed W, along every trajectory, for every t, i and $\lambda > 0$,*

$$\big| h_{t,i} - \operatorname{sign}(W_{\mathrm{in}} x_t)_i \big| \;\le\; 2 \exp\!\big(-2(\lambda |(W_{\mathrm{in}} x_t)_i| - \|W\|_\infty)_+\big), \tag{D.39}$$

with $\|W\|_\infty$ the largest absolute row sum, so as $\lambda \to \infty$ the state tends in every L^p to the memoryless features $\operatorname{sign}(W_{\mathrm{in}} x_t)$. At $d_x = 1$ every unit tends to $\pm \operatorname{sign} x_t$. The fitted readout then tends in L^2 to $\hat{m}\,\sigma^{\top}/(n+\gamma)$, where

$$\sigma = \operatorname{sign} W_{\mathrm{in}}, \qquad \hat{m} = \frac{1}{L} \sum_{t \in I_s} x_{t+s} \operatorname{sign} x_t, \tag{D.40}$$

and its mean and covariance converge as well, at every L.

D.4.2 Proof of Proposition D.8

Proof. We first prove the bound on each unit. Write $z = (W h_{t-1})_i + \lambda (W_{\mathrm{in}} x_t)_i$. Since $h_{t-1} \in [-1,1]^n$, $|(W h_{t-1})_i| \le \|W\|_\infty$. If $\lambda |(W_{\mathrm{in}} x_t)_i| > \|W\|_\infty$, then z has the sign of $(W_{\mathrm{in}} x_t)_i$ and $|z| \ge \lambda |(W_{\mathrm{in}} x_t)_i| - \|W\|_\infty$, so

$$|h_{t,i} - \operatorname{sign}(W_{\mathrm{in}} x_t)_i| = 1 - \tanh|z| = 2/(e^{2|z|} + 1) \le 2 e^{-2|z|}. \tag{D.41}$$

Otherwise the left side is at most 2. This is (D.39), and it holds for every W without a contraction.

The convergence of the state follows from this bound. The right side of (D.39) is at most 2, and it tends to zero as $\lambda \to \infty$ on the event $(W_{\mathrm{in}} x_t)_i \neq 0$, which has probability one. By stationarity its law does not depend on t, so dominated convergence gives $\mathbb{E}|h_{t,i} - \operatorname{sign}(W_{\mathrm{in}} x_t)_i|^p \to 0$ uniformly in t for every p.

We next compute the limit of the readout at $d_x = 1$. Let $\sigma = \operatorname{sign} W_{\mathrm{in}}$. Then $\operatorname{sign}(W_{\mathrm{in}} x_t) = \sigma \operatorname{sign} x_t$ and $(\operatorname{sign} x_t)^2 = 1$ almost surely, so $\hat{\Sigma}_h \to \sigma\sigma^{\top}$ in every L^p, uniformly in L, because it is a window average of bounded terms that converge uniformly in t.

For the cross-moment, $\|x_{t+s}(h_t - \sigma \operatorname{sign} x_t)\|^2 \le 4n\, x_{t+s}^2$, which is integrable by Assumption 1. Dominated convergence therefore gives $\hat{M} \to \hat{m}\,\sigma^\top$ in L^2, uniformly in L by stationarity. The map $(M, \Sigma) \mapsto M(\Sigma + \gamma I)^{-1}$ is Lipschitz in M with constant γ^{-1} and in Σ with constant $\gamma^{-2}\|M\|$ on $\Sigma \succeq 0$. Hence, in L^2,

$$\hat{W} \to \hat{m}\,\sigma^\top(\sigma\sigma^\top + \gamma I)^{-1} = \hat{m}\,\sigma^\top/(n+\gamma), \tag{D.42}$$

where the equality uses $\sigma^\top\sigma = n$, and the mean and covariance at every L converge as well.

In the population,

$$M_y = m_y\sigma^\top, \qquad \Sigma_h = \sigma\sigma^\top, \qquad \bar{W}_y = m_y\sigma^\top/(n+\gamma), \qquad m_y := \mathbb{E}_y[x_{t+s}\operatorname{sign} x_t]. \tag{D.43}$$

With the input standardized to unit variance,

$$R_y^2 = 2\bar{W}_y M_y^\top - \bar{W}_y\Sigma_h\bar{W}_y^\top = m_y^2[2n/(n+\gamma) - n^2/(n+\gamma)^2], \tag{D.44}$$

which tends to m_y^2 as $\gamma/n \to 0$.

For the two-tap pair these quantities are explicit. On the first class $\operatorname{sign} x_t = e_t$ and $m_1 = \mathbb{E}[x_{t+1}e_t] = 1/\sqrt{5}$. On the second class $\operatorname{sign} x_t = e_{t-1}$, which is independent of x_{t+1}, and $m_2 = 0$. The centred terms $x_{t+1}\operatorname{sign} x_t - m_y$ are

$$\frac{2}{\sqrt{5}}e_{t+1}e_t \qquad \text{and} \qquad \frac{2}{\sqrt{5}}(\tfrac{1}{2}e_{t+1} + e_t)e_{t-1}. \tag{D.45}$$

These are products of Rademacher variables that are uncorrelated across lags, with variances $\frac{4}{5}$ and 1. The pooled long-run variance is $\frac{9}{10}$, and the separation per unit length of $\hat{m}$ is $\frac{1}{5}/\frac{9}{10} = \frac{2}{9}$. □

Corollary D.9 (The sampling law at saturation)**.** *Let $d_x = 1$, let every entry of W_{in} be non-zero, let each class be a moving average of iid bounded innovations with $x_t \neq 0$ almost surely, and write*

$$g_t := x_{t+s}\operatorname{sign} x_t, \qquad m_y := \mathbb{E}_y g_t, \tag{D.46}$$

and v_y for the long-run variance of g_t under class y. Take $\lambda \to \infty$ at fixed L and then $L \to \infty$. In this limit $\hat{W}$ is a fixed injective linear image of $\hat{m}$, $\sqrt{L}(\hat{m} - m_y)$ converges in law to $\mathcal{N}(0, v_y)$, and the rule that thresholds $\hat{m}$ at $\frac{1}{2}(m_1 + m_2)$ has error

$$\tfrac{1}{2}\Big[\Phi\big(-\tfrac{1}{2}\sqrt{L}\,|\Delta m|/\sqrt{v_1}\big) + \Phi\big(-\tfrac{1}{2}\sqrt{L}\,|\Delta m|/\sqrt{v_2}\big)\Big] + o(1). \tag{D.47}$$

Proof. By Proposition D.8 the fitted readout tends in L^2 to $\hat{m}\,\sigma^\top/(n+\gamma)$ at every L, a fixed injective linear image of $\hat{m}$. The statistic g_t is a bounded function of finitely many consecutive innovations, so it is a finitely dependent stationary sequence, and the central limit theorem for such sequences gives the stated limit of $\sqrt{L}(\hat{m} - m_y)$. The error of the threshold rule follows. □

The corollary describes the saturated end of the scale axis. Its separation per unit length $(\Delta m)^2/\bar{v}$, with $\bar{v}$ the pooled long-run variance, is $\frac{2}{9}$ on the two-tap pair, 7.4×10^{-3} on the eight-tap pair, computed exactly by enumeration, and about 6.0×10^{-4} on the twin, estimated by simulation. At $L = 256$, 1024 and 4096 the resulting errors are 0.000 on the two-tap pair, 0.246, 0.085 and 0.003 on the eight-tap pair, and 0.42, 0.35 and 0.22 on the twin. In the scale sweep of Appendix F.7 at $\lambda = 100$ the errors are 0.000 on the two-tap pair, 0.005, 0.000 and 0.000 on the eight-tap pair, and 0.40, 0.25 and 0.07 on the twin. The two-tap pair is saturated there, since $|x_t| \ge 0.45$. On the eight-tap pair $|x_t|$ comes as close to zero as 0.004, and on the twin it has a continuous law, so these readouts approach the limit slowly and keep information that saturation discards.

D.5 The lower bound over classifiers

The lower bound of Theorem 4(b) rests on uniform convergence of the laws of the readout over the class of decision regions. For convex regions this is the theorem of Ranga Rao (1962). For regions of bounded algebraic complexity it is the following lemma. For integers $H, \nu \ge 1$, let $\mathcal{A}_{H,\nu}$ be the family of sets

$$\{u \in \mathbb{R}^p : \beta(\mathbf{1}\{p_1(u) > 0\}, \dots, \mathbf{1}\{p_H(u) > 0\}) = 1\}, \tag{D.48}$$

where $p_1, \dots, p_H$ are real polynomials of degree at most ν and $\beta : \{0,1\}^H \to \{0,1\}$ is a Boolean function. A non-strict inequality $p \ge 0$ is the complement of $-p > 0$ and is covered.

Lemma D.10 (Uniform convergence over regions of bounded complexity)**.** *Let $\mu_L \Rightarrow \mu$ be probability laws on $\mathbb{R}^p$, and let μ have a density. Then* $\sup_{A\in\mathcal{A}_{H,\nu}} |\mu_L(A) - \mu(A)| \to 0$.

Proof. Idea. The argument is by contradiction and compactness. After the coefficients are normalized, a sequence of bad regions has a limiting region of the same kind. Skorokhod's representation and the fact that polynomial zero sets are null then make both probabilities converge to the probability of that limit.

Suppose not. Then there are $\varepsilon > 0$, a subsequence L_j and sets $A_j \in \mathcal{A}_{H,\nu}$ with $|\mu_{L_j}(A_j) - \mu(A_j)| > \varepsilon$. There are finitely many Boolean functions, so along a further subsequence all A_j use one β, and the set of indices h for which $p_{h,j}$ is the zero polynomial does not depend on j.

For these h the atom $\{p_{h,j} > 0\}$ is empty. For the other h, dividing $p_{h,j}$ by the Euclidean norm of its coefficient vector does not change $\{p_{h,j} > 0\}$, and by compactness of the unit sphere a further subsequence has $p_{h,j} \to p_h$ coefficientwise, with p_h a polynomial of degree at most ν with unit coefficient norm, hence not zero. Let A be the set defined by β and the limits p_h.

By Skorokhod's representation there are $U_j \sim \mu_{L_j}$ and $U \sim \mu$ on one probability space with $U_j \to U$ almost surely. Then $p_{h,j}(U_j) \to p_h(U)$ almost surely. The zero set of a non-zero polynomial is Lebesgue-null and μ has a density, so $p_h(U) \neq 0$ almost surely, and $\mathbf{1}\{p_{h,j}(U_j) > 0\} \to \mathbf{1}\{p_h(U) > 0\}$ almost surely for every h. Hence $\mathbf{1}_{A_j}(U_j) \to \mathbf{1}_A(U)$ almost surely, and dominated convergence gives $\mu_{L_j}(A_j) \to \mu(A)$. The same argument with U in place of U_j gives $\mu(A_j) \to \mu(A)$, which contradicts $|\mu_{L_j}(A_j) - \mu(A_j)| > \varepsilon$. □

The convex case covers every linear classifier of the readout, the Euclidean nearest-centroid rule for two classes, and a linear classifier applied after an affine map and L2 normalization, with or without centring. For the last case, the region $\{w : a^\top A(w - c) > b\,\|A(w - c)\|\}$ is a convex cone when $b \geq 0$ and has a convex complement when $b < 0$.

The case $\mathcal{A}_{H,\nu}$ covers quadratic discriminant analysis ($H = 1$, $\nu = 2$), decision trees and forests with a bounded number of nodes (axis-aligned half-spaces combined by a Boolean function), and ReLU networks of a fixed architecture, whose decision region is a Boolean combination of a bounded number of half-spaces. It also covers the k-nearest-neighbour rule with a fixed number of training sequences, whose region is decided by the signs of the affine functions $\|u - x_i\|^2 - \|u - x_{i'}\|^2$. Each of these remains in some $\mathcal{A}_{H,\nu}$ after an affine map and L2 normalization. With $\tilde{u}$ affine in u, $r := \|\tilde{u}\|$ and p a polynomial of degree ν,

$$r^\nu p(\tilde{u}/r) = E(\tilde{u}) + r\,O(\tilde{u}) \tag{D.49}$$

with polynomials E, O of degree at most ν. With $Q := E^2 - r^2O^2$, on $\{r > 0\}$

$$\{E+rO > 0\} = \big(\{E > 0\}\cap(\{O \geq 0\}\cup\{Q > 0\})\big)\cup\big(\{O > 0\}\cap(\{E \geq 0\}\cup\{Q < 0\})\big), \tag{D.50}$$

which is decided by the signs of $\pm E$, $\pm O$ and $\pm Q$. Together with $\{r^2 > 0\}$, and a fixed label at $\tilde{u} = 0$, a rule in $\mathcal{A}_{H,\nu}$ therefore becomes one in $\mathcal{A}_{6H+1,\,2\nu+2}$.

The complexity must not grow with L. The argument does not cover rules whose complexity grows with L, which may exploit the finer structure of the readout when the innovations take finitely many values and its law has finite support.

D.6 Proof of Theorem 4 and an exact check at sixth order

Proof. Idea. The proof repeats that of Proposition D.2 at order $2k + 1$. Expanding $\tanh$ to degree $2k + 1$ with a pathwise remainder turns every statistic into a window average that does not depend on λ. The noise is still that of the linear term, at order $\lambda/\sqrt{L}$, while the class means first differ at order λ^{2k+1}, and the scaling $\lambda^{4k}L \to \zeta$ makes the two comparable.

We start with the pathwise expansion. Let T_{2k+1} be the Taylor polynomial of $\tanh$ of degree $2k+1$, so that $|\tanh z - T_{2k+1}(z)| \leq C|z|^{2k+3}$ for $|z| \leq 1$. Define $h_t^{(2j+1)}$ for $j \leq k$ by the recursions (C.25) from zero initial states. As in part (a) of the proof of Proposition D.2, $\sup_t \|h_t\|_D \leq \lambda C_0$, and the remainder satisfies

$$r_t := h_t - \sum_{j\leq k} \lambda^{2j+1} h_t^{(2j+1)}, \qquad \|r_t\|_D \leq q(1 + C\lambda^2)\|r_{t-1}\|_D + C\lambda^{2k+3}, \tag{D.51}$$

so $\sup_t \|r_t\|_D \le C\lambda^{2k+3}$ uniformly in L and along every trajectory. Hence

$$\hat{M} = \sum_{j\le k} \lambda^{2j+1}\hat{M}_{2j+1} + O(\lambda^{2k+3}), \qquad \hat{\Sigma}_h = \sum_{j\le k} \lambda^{2j+2}\hat{\Sigma}_{(2j+2)} + O(\lambda^{2k+4}), \tag{D.52}$$

where $\hat{M}_{2j+1}$ and $\hat{\Sigma}_{(2j+2)}$ are window averages of products of x_{t+s} and the coefficients $h_t^{(2a+1)}$. These statistics do not depend on λ.

We next control the fluctuations of these statistics. Each $h_t^{(2a+1)}$ is a geometrically weighted sum over lags of polynomials of degree $2a+1$ in the inputs, so the truncation argument of Lemma D.1(ii) applies to it. Truncating every Neumann series in (C.26) at lag J changes $h_t^{(2a+1)}$ by $O((J+1)^{2a}q^J)$, by induction on a through the recursion (C.25), and the choice of J in the proof of Lemma D.1(ii) then gives each entry of these statistics a dependence measure $\delta_2(j) \le C(j+1)^{2k}q^{j/2}$, which is summable. From $h_0 = 0$ the same induction gives transients $O((t+1)^{2a}q^t)$ for $h^{(2a+1)}$, which are summable. By Lemma D.1(iii) their means are the stationary values up to $O(1/L)$, and their fluctuations are $O_P(L^{-1/2})$, uniformly in λ.

The readout is $\hat{W} = \hat{M}(\hat{\Sigma}_h + \gamma I)^{-1}$, a smooth function of these statistics whose derivatives are bounded because $\gamma > 0$. In its first-order expansion around the class means of the statistics, only the fluctuation of $\hat{M}_1$ contributes at order $\lambda/\sqrt{L}$. The fluctuations of $\hat{M}_{2j+1}$ with $j \ge 1$ enter with λ^{2j+1}, those of $\hat{\Sigma}_h$ with λ^3, and the second-order terms and the difference between $\mathbb{E}_y\hat{W}$ and the readout at the class means of the statistics are $O(\lambda/L)$. After multiplication by $\sqrt{L}/\lambda$ these terms are $O_P(\lambda^2) + O(L^{-1/2})$, and the remainder $O(\lambda^{2k+3})$ becomes $O(\sqrt{L\lambda^{4k}}\,\lambda^2) = o(1)$. All of them vanish in the limit. Hence on class y,

$$\sqrt{L}(\mathrm{vec}\hat{W} - \mathbb{E}_y\mathrm{vec}\hat{W})/\lambda = (R_\lambda \otimes I)\xi_L + o_P(1), \tag{D.53}$$

with $\xi_L \Rightarrow \mathcal{N}(0, V_1)$, by Lemma D.1(iii) applied to $\hat{M}_1$.

We now turn to the class means. The class means of the statistics are the stationary values up to $O(1/L)$, and the readout evaluated at the stationary values is the population readout of the truncated expansion, which differs from $\bar{W}_{g_s,y}$ by $O(\lambda^{2k+3})$. Hence, by Theorem 3(a),

$$\Delta\mathbb{E}\hat{W} = \Delta\bar{W} + O(\lambda/L) + O(\lambda^{2k+3}) = \gamma^{-1}\lambda^{2k+1}\Delta M_{g_s}^{(2k+1)} + O(\lambda^{2k+3}) + O(\lambda/L). \tag{D.54}$$

After multiplication by $\sqrt{L}/\lambda$ this is $\gamma^{-1}\sqrt{\lambda^{4k}L}\,\delta_k + o(1) \to \gamma^{-1}\sqrt{\zeta}\,\delta_k$.

With $R_\lambda \to \gamma^{-1}I$, the centred and scaled readout $u := \sqrt{L}(\mathrm{vec}\hat{W} - \langle\hat{W}\rangle_L)/\lambda$ converges in law on class y to

$$Q_y := \mathcal{N}\big(\pm\tfrac{1}{2}\sqrt{\zeta}\,\gamma^{-1}\delta_k,\ \gamma^{-2}V_1\big), \tag{D.55}$$

which is part (a).

For part (b), both limits have densities because $V_1 \succ 0$. By the theorem of Ranga Rao (1962), the probability that u falls in a convex Borel set therefore converges to its Q_y-probability uniformly over all such sets, and by Lemma D.10 the same holds uniformly over $\mathcal{A}_{H,\nu}$. The map from $\mathrm{vec}\hat{W}$ to u is affine, so it preserves convexity and the degree of every polynomial. Let the rule assign class 1 on a region A that is convex in u or lies in $\mathcal{A}_{H,\nu}$. The case in which the region of class 2 is convex is the same, with A replaced by its complement. Conditionally on the training data, the test error is

$$\frac{1}{2}P_1(u \notin A) + \frac{1}{2}P_2(u \in A) = \frac{1}{2}Q_1(A^c) + \frac{1}{2}Q_2(A) + o(1), \tag{D.56}$$

uniformly in A, and hence also after averaging over the training data. The right side is at least the Bayes error of Q_1 against Q_2. These two laws share their covariance, and the squared Mahalanobis distance between their means is $\zeta\,\delta_k^\top V_1^{-1}\delta_k = \zeta\kappa_k$, so the Bayes error is $\Phi(-\frac{1}{2}\sqrt{\zeta\kappa_k})$. It is attained by the half-space of LDA with the population parameters.

Finally, for $k \ge 2$ the shared linear noise is automatic, as the statement says. The matrix V_1 is the long-run covariance of $\hat{M}_1$, a combination of the sample autocovariances, and by (D.18) and its general form it is a function of the joint moments of order at most four. For $k \ge 2$ the classes share these. At $k = 1$ scalar reversal pairs share V_1 as well. The sample autocovariances of a path and of its reversal differ only by edge terms whose mean absolute value is $O(1/L)$ (proof of Lemma C.2), which vanish after the $\sqrt{L}$ scaling, and the reversed path of one class is a path of the other. □

The sixth-order pair of Appendix F.9 is the two-tap reversal pair with a symmetric four-point innovation law of fourth moment three. It has vanishing joint cumulants of orders three to five, so it shares every joint moment of order at most five. It differs in

$$c_2(1) = \mathrm{cum}_6(e)(\tfrac{1}{2} - \tfrac{1}{32})/(\tfrac{5}{4})^3, \tag{D.57}$$

where $\mathrm{cum}_6(e)$ is the sixth cumulant of the innovations. It is therefore an instance of Theorem 4 with $k = 2$ for generic reservoirs, by Theorem 3(b), and its separation grows as $\lambda^8 L$.

At sixth order the theorem can be checked by exact computation, without simulation, on a pair of this kind. It makes three claims at once: a mean difference of order λ^{2k+1}, a noise of order $\lambda/\sqrt{L}$, and their ratio κ_k.

The innovations take the three-point Gauss–Hermite law, $\pm\sqrt{3}$ with probability $\frac{1}{6}$ each and 0 with $\frac{2}{3}$, which matches the Gaussian moments up to order five and has sixth cumulant -6. The pair is $x_t = (e_t + \frac{1}{2}e_{t-1})/\sqrt{5/4}$ against its reversal, fed to one memoryless unit with input weight 1 at $\gamma = 1$ and the next-step target. Then

$$\delta_2 = -\frac{24}{125}, \qquad V_1 = \frac{37}{25}, \qquad \kappa_2 = \frac{576}{23125} \approx 0.024908. \tag{D.58}$$

For the full $\tanh$ readout we compute the population readouts and the exact long-run variance of the fitted readout from its influence function, which depends on three consecutive innovations, by enumeration at 60 digits. The separation per unit length divided by λ^8 is 0.024906, 0.024900, 0.024875 and 0.024777 at $\lambda = 0.005$, 0.01, 0.02 and 0.04, and its log-log slope between consecutive scales is 7.9996, 7.9986 and 7.9943.

Lemma D.11 (Ridge propagation of leading-order separation)**.** *Fix $\gamma > 0$. Suppose two stationary processes P_1, P_2 generate population cross-moments and reservoir covariances admitting the small-λ scaling*

$$M_i(\lambda) = O(\lambda), \qquad \Sigma_{h,i}(\lambda) = O(\lambda^2), \qquad i = 1, 2, \tag{D.59}$$

with leading class difference at Taylor order $r \geq 1$:

$$\Delta M(\lambda) \;=\; \lambda^r A + O(\lambda^{r+2}), \qquad A \in \mathbb{R}^{d_x \times n}, \tag{D.60}$$

and $\Delta\Sigma_h(\lambda) = O(\lambda^{r+1})$. Then the population fitted readout difference satisfies

$$\Delta\bar{W}(\lambda) \;:=\; \bar{W}_1(\lambda) - \bar{W}_2(\lambda) \;=\; \gamma^{-1}\lambda^r A + O(\lambda^{r+2}). \tag{D.61}$$

The covariance condition is needed (Remark D.12).

Proof. Write $A_i = \Sigma_{h,i}(\lambda) + \gamma I$, so $A_i \succeq \gamma I$ and $\|A_i^{-1}\|_{\mathrm{op}} \leq \gamma^{-1}$ at every λ and every $\gamma > 0$. The exact resolvent identity

$$A_1^{-1} - A_2^{-1} = -A_1^{-1}(A_1 - A_2)A_2^{-1}, \tag{D.62}$$

applied to $\bar{W}_i = M_i A_i^{-1}$, gives

$$\Delta\bar{W}(\lambda) \;=\; (\Delta M)\, A_1^{-1} \;-\; M_2\, A_1^{-1}(\Delta\Sigma_h)\, A_2^{-1}. \tag{D.63}$$

This is (A.3) for the pair at input scale λ, and it involves no expansion of either inverse.

For the first term, $A_i^{-1} = \gamma^{-1}(I - A_i^{-1}\Sigma_{h,i})$ gives $\|A_i^{-1} - \gamma^{-1}I\|_{\mathrm{op}} \leq \gamma^{-2}\|\Sigma_{h,i}\|_{\mathrm{op}} = O(\lambda^2)$, so

$$(\Delta M)A_1^{-1} = \gamma^{-1}\Delta M + O(\lambda^r)\, O(\lambda^2) = \gamma^{-1}\lambda^r A + O(\lambda^{r+2}). \tag{D.64}$$

For the second,

$$\|M_2 A_1^{-1}(\Delta\Sigma_h)A_2^{-1}\|_F \leq \gamma^{-2}\|M_2\|_F\|\Delta\Sigma_h\|_{\mathrm{op}} = O(\lambda)\, O(\lambda^{r+1}) = O(\lambda^{r+2}). \tag{D.65}$$

Adding the two gives (D.61). □

Remark D.12 (The covariance condition is needed)**.** *Without $\Delta\Sigma_h = O(\lambda^{r+1})$ the readout does not inherit the cross-moment's coefficient. Take $n = d_x = 1$, $W = 0$, $W_{\mathrm{in}} = \gamma = 1$, $s = 1$, and both classes $x_t = e_t + 2e_{t-1} - \frac{1}{4}e_{t-2}$, with Rademacher innovations for one and uniform innovations on $[-\sqrt{3}, \sqrt{3}]$ for the other. They share every autocovariance and $\Delta\mathbb{E}[x_{t+1}x_t^3] = 0$, so the cross-moment first differs at $r = 5$, with $\Delta M^{(5)} = -9371/448$. On the other hand, $\Delta\mathbb{E}x_t^4 \neq 0$ gives $\Delta\Sigma_h^{(4)} = 1451/160$. Then*

$$\Delta\bar{W}(\lambda) = \lambda^5\big(\Delta M^{(5)} - M^{(1)}\Delta\Sigma_h^{(4)}\big) + O(\lambda^7) = -\tfrac{38663}{1120}\lambda^5 + O(\lambda^7), \tag{D.66}$$

which is 1.65 times the inherited coefficient $\gamma^{-1}\Delta M^{(5)}$. Second-order-matched pairs at $r = 3$ satisfy the condition, since their $\Delta\Sigma_h$ is $O(\lambda^4)$.

E THE EXACT SPLIT, SIGN SYMMETRY AND THE LAG ENVELOPE

This appendix gives the exact split of the state into a linear and a higher-order part, which §5.1 uses as control variates, and proves that the weight of the linearized and cubic statistics decays geometrically in the lag.

E.1 THE SPLIT

Part (a) of Proposition D.2 has an exact counterpart that holds at every length and every input scale. Let $a_t := h_t^{(1)}$ be the linearized state and set, entrywise,

$$\psi_\lambda(v) := (\tanh(\lambda v) - \lambda v)/\lambda^3, \qquad \psi_0(v) := -v^{\odot 3}/3. \tag{E.1}$$

Define, from zero initial states,

$$a_t = W a_{t-1} + W_{\mathrm{in}} x_t, \qquad v_t = a_t + \lambda^2 W b_{t-1}, \qquad b_t = W b_{t-1} + \psi_\lambda(v_t). \tag{E.2}$$

Since the pre-activation is λv_t and $\tanh(\lambda v) = \lambda v + \lambda^3 \psi_\lambda(v)$, the state and every fitted cross-moment split exactly, at every t, every scale and every length. We call this decomposition the *split*:

$$h_t = \lambda a_t + \lambda^3 b_t, \qquad \hat{M} = \lambda \hat{M}_1 + \lambda^3 \hat{M}_{3,\lambda}, \qquad \hat{M}_{3,\lambda} = \tfrac{1}{L} \sum_{t \in I_s} x_{t+s} b_t^\top. \tag{E.3}$$

At $\lambda = 0$, $b_t = c_t := W c_{t-1} - a_t^{\odot 3}/3$ is the cubic state $h^{(3)}$ of Lemma C.5. The readout splits in the same way as the cross-moment. Its higher-order part is $\lambda^3 \hat{M}_{3,\lambda}(\hat{\Sigma}_h + \gamma I)^{-1}$ at every λ and every ridge, including the selected scales, and it is computed by one additional recurrence on the same sequence.

The split provides the control variates with which §5.1 measures the entry orders. Its zero-scale limit

$$\hat{M}_{3,0} = \frac{1}{L} \sum x_{t+s} c_t^\top \tag{E.4}$$

is computed in one pass alongside $\hat{M}_1$.

Remark E.1 (Which reservoirs satisfy Assumption 2)**.** *Assumption 2 holds with $D = I$ and $q = \rho$ for $W = \rho Q$ with Q orthogonal, and at the parameters below for the cycle-with-jumps topology. We searched over positive diagonal D at $n = 100$, $\rho = 0.9$ and three seeds. The smallest $\|D^{1/2} W D^{-1/2}\|_2$ found is 0.900 for the orthogonal construction and 0.946 for cycle-with-jumps, which certifies both. It is 1.504 for a dense and 1.477 for a sparse (0.1) Gaussian W rescaled to $\rho(W) = 0.9$, and for these the search gives no certificate. Trajectories on the Gaussian reservoirs forget their initial state at rate $\approx \rho(W)$ in every run, and the experiments report them alongside the orthogonal arm.*

Remark E.2 (Exact sign invariance of the pipeline)**.** *With $h_0 = 0$ and no bias, the oddness of $\tanh$ gives $h_t(-x) = -h_t(x)$ for every t by induction. Hence the state covariance, the target–state cross-moment, the fitted readout at any ridge, and every deterministic post-processing of them coincide for x and $-x$ at every T. A process and its reflection are therefore indistinguishable to the empirical pipeline. For example, the iid law that takes the value 2 with probability $\frac{1}{3}$ and -1 with probability $\frac{2}{3}$ and its reflection share mean and variance and have third moments ± 2. They induce the same law of the representation, so every classifier on it has Bayes error $\frac{1}{2}$. For the same reason $\hat{W}$ is odd in λ, and its coefficient of λ^{2k+1} is a polynomial of degree $2k+2$ in the input path. Odd cumulants still enter through even products, as in $\mu_6 = \mathrm{cum}_6 + 15\mathrm{cum}_2\mathrm{cum}_4 + 10\mathrm{cum}_3^2 + 15\mathrm{cum}_2^3$.*

E.2 THE LAG ENVELOPE

Let $w(\ell) := \|W^\ell W_{\mathrm{in}}\|_2^2$ be the squared weight that the linearized state places on $x_{t-\ell}$, and let $u := \|W_{\mathrm{in}}\|_2$. At $d_x = 1$, for a lag triple $j_1 \le j_2 \le j_3$, write

$$\begin{aligned} \hat{m}(s; j_1, j_2, j_3) &:= \frac{1}{L} \sum_{t \in I_s} x_{t+s} x_{t-j_1} x_{t-j_2} x_{t-j_3}, \\ G(j_1, j_2, j_3) &:= -\frac{\mu}{3} \sum_{j=0}^{j_1} W^j \bigodot_{k=1}^{3} W^{j_k - j} W_{\mathrm{in}}, \end{aligned} \tag{E.5}$$

where μ is the number of distinct orderings of (j_1, j_2, j_3). Expanding $c_t = -\frac{1}{3}\sum_{j\geq 0} W^j a_{t-j}^{\odot 3}$ in the lags gives

$$\hat{M}_{3,0} = \sum_{j_1 \leq j_2 \leq j_3} \hat{m}(s; j_1, j_2, j_3)\, G(j_1, j_2, j_3)^\top, \tag{E.6}$$

up to the lags that precede the first observation, so $G(j_1, j_2, j_3)$ is the coefficient of $\hat{m}(s; j_1, j_2, j_3)$ in the cubic statistic.

Proposition E.3 (The lag envelope)**.** (a) *If $\|W\|_2 \leq q$ then $w(\ell+1) \leq q^2 w(\ell)$ for every ℓ, so for $q \leq 1$ the profile is non-increasing. For $q < 1$ the share of the state's input weight at lags $\geq \ell$ is at most $q^{2\ell}/(1-q^2)$, and it is exactly $\rho^{2\ell}$ when $W = \rho Q$ with $Q^\top Q = I$.* (b) *Hence no reservoir with $\|W\|_2 \leq 1$ places its weight in a band. A profile that peaks at a non-zero lag requires $\|W\|_2 > 1$, which $\rho(W) < 1$ permits.* (c) *The cubic coefficient obeys the same envelope. Under $\|W\|_2 \leq q < 1$, $\|G(j_1, j_2, j_3)\|_2 = O(q^{\,j_2+j_3-j_1})$, and hence $O(q^{j_3})$, with the explicit constant of* (E.9)*. Truncating the statistic to triples with $j_3 < J$ therefore changes it by $O(Jq^J)$.*

Proof. (a) $\|W^{\ell+1} W_{\text{in}}\|_2 \leq \|W\|_2 \|W^\ell W_{\text{in}}\|_2$ gives $w(\ell+1) \leq q^2 w(\ell)$ and, iterated, $w(\ell+i) \leq q^{2i} w(\ell)$. Hence

$$\sum_{k\geq \ell} w(k) \leq w(\ell)/(1-q^2) \leq q^{2\ell} w(0)/(1-q^2), \qquad \sum_{k\geq 0} w(k) \geq w(0), \tag{E.7}$$

which is the stated share. For $W = \rho Q$ with Q orthogonal, $\|W^\ell W_{\text{in}}\|_2 = \rho^\ell u$ because Q is an isometry, so $w(\ell) = \rho^{2\ell} u^2$, both sums are geometric and the share is $\rho^{2\ell}$ exactly.

(b) Part (a) at $q = 1$ makes w non-increasing, so a peak at a non-zero lag forces $\|W\|_2 > 1$, which is compatible with $\rho(W) < 1$. For example, the 2×2 nilpotent W with $W_{12} = \beta$ and $W_{\text{in}} = e_2$ has $\rho(W) = 0$ and $w = (1, \beta^2, 0, \ldots)$, which peaks at lag one once $\beta > 1$.

(c) Fix a triple $j_1 \leq j_2 \leq j_3$ and write $p_j := \bigodot_{k=1}^3 W^{j_k - j} W_{\text{in}}$ for the elementwise product, so that $G(j_1, j_2, j_3) = -\frac{\mu}{3}\sum_{j=0}^{j_1} W^j p_j$. Each factor obeys $\|W^{j_k-j} W_{\text{in}}\|_\infty \leq \|W^{j_k-j} W_{\text{in}}\|_2 \leq q^{j_k-j} u$, so

$$\|p_j\|_2 \leq \sqrt{n}\|p_j\|_\infty \leq \sqrt{n}\, u^3 q^{\sum_k j_k - 3j}, \qquad \|W^j p_j\|_2 \leq q^j \|p_j\|_2. \tag{E.8}$$

Summing and using $\sum_{j=0}^{j_1} q^{-2j} \leq q^{-2j_1}/(1-q^2)$,

$$\|G(j_1, j_2, j_3)\|_2 \;\leq\; \frac{\mu\sqrt{n}\, u^3}{3(1-q^2)}\, q^{\sum_k j_k - 2j_1} \;=\; \frac{\mu\sqrt{n}\, u^3}{3(1-q^2)}\, q^{\,j_2+j_3-j_1}, \tag{E.9}$$

and $j_2 + j_3 - j_1 \geq j_3$ gives the second form.

For the truncation, the triples with $j_3 = \ell$ have envelope $q^{\ell + j_2 - j_1}$, and its sum over $j_1 \leq j_2 \leq \ell$ is at most $(\ell+1)q^\ell/(1-q)$. Summing over $\ell \geq J$ against a bounded moment array gives $O(Jq^J)$. □

Remark E.4 (The cubic statistic as a truncated polynomial mean embedding)**.** *Write $\hat{m}$ for the sample fourth-order lag moments*

$$\frac{1}{L}\sum_t x_{t+s} x_{t-j_1} x_{t-j_2} x_{t-j_3} \tag{E.10}$$

over all lag triples, so that $\hat{M}_{3,0} = \mathcal{K}\hat{m}$ where $\mathcal{K}$ is the fixed linear map whose weights are bounded by part (c) above. Truncating the triples at lag J and writing $u_t = (x_t, \ldots, x_{t-J+1})$ and $\phi(y, u) = y\, u^{\otimes 3}$,

$$\hat{M}_{3,0} \;=\; \mathcal{K}_J \hat{\mu}_\phi + R_J, \qquad \hat{\mu}_\phi := \tfrac{1}{L}\sum_t \phi(x_{t+s}, u_t), \qquad \|R_J\| = O(Jq^J), \tag{E.11}$$

where the bound on the remainder follows from part (c).

Up to the remainder, $\hat{M}_{3,0}$ is a fixed finite-rank image of the mean embedding $\hat{\mu}_\phi$ whose kernel is $k((y,u),(y',u')) = yy'(u^\top u')^3$, once the repeated monomials are given their multinomial weights. The delay-polynomial features of Gauthier et al. (2021) are the unweighted coordinates of the truncation together with a linear block, and they span the same space under a different inner product.

F EXPERIMENTAL PROTOCOLS AND SUPPLEMENTARY RESULTS

F.1 COMMON CONVENTIONS

Every experiment runs the $\tanh$ recurrence of §2 from $h_0 = 0$ in float64 and computes the ridge readout in closed form from the normal equations. Where the linearized and cubic states of (E.2) are used, they are driven on the same sequence. Each subsection states its own width, spectral radius, reservoir family, input-weight law, washout, input scale, ridge parameterization (the per-sample γ of (2), or the relative level g with $\gamma = g\,\mathrm{tr}(\hat{\Sigma}_h)/n$), sequence counts, lengths and seeds. The unit of replication is an independent draw of reservoir and data. Where sub-samples of one draw are averaged, the text says so. Every data-dependent transformation (standardization, shrinkage or ridge selection) is fitted on training sequences only.

Shrinkage LDA is LDA with Ledoit–Wolf shrinkage estimated on the training set. In each class, each coordinate is standardized by its class sample standard deviation, the shrinkage intensity is computed on the standardized sample and applied toward the scaled identity, and the shrunk covariance is rescaled. The two class covariances are then averaged with their frequencies, and the discriminant is the inverse of this covariance applied to the mean difference. The features enter it unscaled.

Two bounded pairs, the two-tap pair and the eight-tap pair, are used at the next-step target. In each pair one class is the time reversal of the other, and $e_t = \pm 1$ with equal probability. The two-tap pair compares

$$x_t = (e_t + \vartheta_1 e_{t-1})/\sqrt{1+\vartheta_1^2} \quad \text{with} \quad (\vartheta_1 e_t + e_{t-1})/\sqrt{1+\vartheta_1^2}, \tag{F.1}$$

where $\vartheta_1 = \frac{1}{2}$. The two classes share one autocovariance ($\Gamma(1) = \vartheta_1/(1+\vartheta_1^2)$, and zero beyond lag one) and one marginal law, since the coefficients of two iid symmetric innovations are exchanged. Their first difference is at the lag-one fourth-order joint cumulant,

$$\Delta\mathrm{cum}(x_{t+1}, x_t, x_t, x_t) = \mathrm{cum}_4(e)(\vartheta_1 - \vartheta_1^3)/(1+\vartheta_1^2)^2 = -0.48, \tag{F.2}$$

with $\mathrm{cum}_4(e) = -2$. The input bound is $B_x = 1.34$.

The eight-tap pair compares

$$x_t = \sum_{j=0}^{7} \vartheta_j e_{t-j}/\|\vartheta\|_2, \qquad \vartheta_j = 0.85^j \cos(2\pi \cdot 0.11 j), \tag{F.3}$$

against the reversed vector. These are the first eight coefficients of the twin, defined below. The same argument gives the same invariants. Here $\Delta\mathrm{cum}(x_{t+1}, x_t, x_t, x_t) = -0.21$, and a difference is also spread over the fourth-order lag triples $(x_{t+1}, x_{t-j_1}, x_{t-j_2}, x_{t-j_3})$. The input bound is

$$B_x = \sum_j |\vartheta_j|/\|\vartheta\|_2 = 2.29. \tag{F.4}$$

The twin is a third reversal pair. Both classes are moving averages

$$x_t = \sum_{j=0}^{15} \vartheta_j e_{t-j} \tag{F.5}$$

of iid standardized Laplace innovations ($\mathrm{cum}_4 = 3$), class 1 through $\vartheta_j = 0.85^{\,j} \cos(2\pi \cdot 0.11\, j)$ and class 2 through the reversed vector. Each realization is standardized to zero mean and unit variance, except where a subsection divides the coefficients by the population standard deviation instead. The two autocovariance sequences agree to 1.1×10^{-16} at every lag. The reversal-sensitive coefficient $\sum_j \vartheta_j \vartheta_{j+1}^3$ differs between the classes by a factor 2.16.

F.2 RAW SEQUENCES AND FITTED READOUTS

Rules on the raw sequences stay at chance, while the fitted readouts separate the classes as the length grows (§2).

Protocol. The pairs are the three reversal pairs, the moving averages $x_t \propto e_t + \vartheta_1 e_{t-1}$ with $\vartheta_1 \in \{0.8, 0.3\}$ and Gaussian innovations, and iid Gaussian against iid standardized Laplace input,

all with unit variance. One reservoir is used, with $n = 20$, $W = 0.9Q$, a unit-norm Gaussian W_{in}, $\lambda = 1$, washout 50 and the relative ridge $g = 10^{-3}$. For each $L \in \{64, 256, 1024, 4096\}$ we draw 625 sequences per class. The training sets are five disjoint blocks of $N \in \{1, 5, 25\}$ sequences per class, and the last 500 sequences per class form the test set. The rules are 1-NN, 5-NN, an RBF support vector machine with $C = 1$ and bandwidth equal to the median squared distance of the training set, a random forest with 200 trees, and shrinkage LDA. Readouts are standardized on the training set, except at $N = 1$, and raw sequences are used as they are.

At $N = 1$ the 1-NN limit of Proposition B.1(a) is exactly $\frac{1}{2}$. At larger N we compute it by simulating the Gaussian scores of its proof, with 2×10^5 draws.

Table 4: Raw sequences and fitted readouts at $L = 4096$ (§2), test error, mean over five training sets. *Raw*: the best of 1-NN, 5-NN, RBF-SVM and shrinkage LDA, and the random forest, with 25 training sequences per class. *Readouts*: 1-NN with one training sequence per class, and shrinkage LDA with 25.

	Raw, $N = 25$		Next-step readout		Self-step readout	
Pair	distance, LDA	forest	1-NN, N=1	LDA, N=25	1-NN, N=1	LDA, N=25
Two-tap	0.484	0.490	0.003	0.001	0.000	0.000
Eight-tap	0.497	0.513	0.113	0.029	0.007	0.000
Twin	0.491	0.511	0.166	0.070	0.291	0.018
Moving averages	0.487	0.501	0.000	0.000	0.000	0.000
iid Gaussian–Laplace	0.497	0.307	0.502	0.496	0.000	0.000

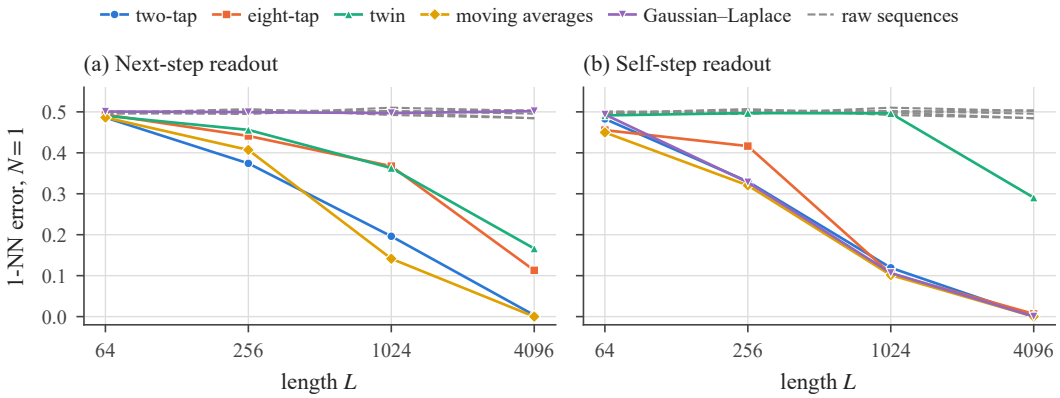


Figure 4: Raw sequences and fitted readouts (§2): 1-NN test error with one training sequence per class against the length, on one reservoir ($n = 20$, $\rho = 0.9$, $\lambda = 1$). Dashed grey: the raw sequences, all five pairs. Coloured: the fitted readout at the next-step target (a) and at the self-step target (b). At the next-step target the Gaussian–Laplace pair has no mean signal, since both conditional means vanish (Corollary 1).

Results. Figure 4 gives 1-NN with one training sequence per class at every length, and Table 4 gives all five rules at $L = 4096$. Over all pairs, lengths, rules other than the random forest and training sizes, the error on the raw sequences lies between 0.484 and 0.516. The random forest on the raw coordinates is between 0.452 and 0.516 on the pairs other than the iid pair. On the iid pair, with 25 sequences per class, it reaches 0.43, 0.35, 0.33 and 0.31 at $L = 64$, 256, 1024 and 4096.

F.3 Families of dynamical laws

This subsection gives the protocol and the results for the families of dynamical laws of §5.2. Every sequence is standardized, so the variance carries no class information, and each class is a mixture of stationary laws.

In every class, each sequence draws its parameters independently, with e_t iid standard Gaussian. The autoregressive class is $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + e_t$ with $\phi_1 \sim U(0.3, 0.7)$ and $\phi_2 \sim U(-0.3, 0)$. The

threshold class is $x_t = a_1 x_{t-1} + e_t$ if $x_{t-1} < 0$ and $a_2 x_{t-1} + e_t$ otherwise, with $a_1 \sim U(0.5, 0.8)$ and $a_2 \sim U(-0.1, 0.3)$. The bilinear class is $x_t = \phi x_{t-1} + b\, x_{t-1} e_{t-1} + e_t$ with $\phi \sim U(0.3, 0.6)$ and $b \sim U(0.3, 0.5)$. The AR-GARCH class is

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + u_t, \qquad u_t = s_t e_t, \qquad s_t^2 = \omega + \alpha u_{t-1}^2 + \beta s_{t-1}^2, \tag{F.6}$$

with (ϕ_1, ϕ_2) drawn as in the autoregressive class, so that the two classes share the law of the autoregressive coefficients, and with $\alpha \sim U(0.1, 0.2)$, $\beta \sim U(0.6, 0.75)$ and $\omega = 1 - \alpha - \beta$. The first 200 steps are discarded.

The oscillator task has three classes of two Kuramoto oscillators with coupling 0, 0.3 and 1, observed through $(\sin\theta_1, \sin\theta_2)$, with random natural frequencies and initial phases. Every sequence is standardized to zero mean and unit variance in each channel.

Each sequence is then mapped to several representations. The generic statistics of a scalar sequence are the autocorrelations at lags 1 to 8, the skewness and kurtosis, $\overline{x_t x_{t-k}^2}$ for $k \le 3$, $\overline{x_t^2 x_{t-k}^2}$ for $k \le 4$ and $\overline{x_t^3 x_{t-k}}$ for $k \le 3$, 20 in all. For two channels they are the autocorrelations at lags 1 to 5 and the kurtosis of each channel, the cross-correlations at lags 0, ± 1 and ± 2, and four fourth-order cross moments, 21 in all. The AR coefficients are least-squares fits of order 20, or a vector AR of order 10 with 40 coefficients for two channels. The readout uses $W = 0.9Q$, one unit-norm Gaussian input column per channel, the next-step target, washout 100 and the relative ridge $g = 10^{-3}$, with $n = 20$ for one channel and $n = 10$ for two, so that it has 20 coordinates.

The second readout adds a bias with iid $\mathcal{N}(0, \frac{1}{4})$ entries inside the $\tanh$, fits with an intercept, and adds the target $x_{t+1}^2 - 1$ for each channel, with half the units so that its dimension is again 20. For every readout the scale is chosen among $\{0.1, 0.3, 1, 3\}$ by the pooled held-out one-step R^2 on the training sequences, averaged over the targets, fitted on the first three quarters of each window and scored on the rest, without labels. The squared target enters this score, because in a class of martingale differences the R^2 of the first target is zero at every scale. Features other than the raw sequence are standardized.

Protocol. We use one set of sequences, generated with a fixed seed for each task, four reservoir draws, five disjoint training sets of 50 sequences per class, and 200 test sequences per class, at $L = 1024$. The five training sets share the test set, so the spread over training sets is not that of independent repetitions. Rules are 1-NN, an RBF support vector machine with $C = 1$ and bandwidth the median squared distance, and shrinkage LDA.

Results. Table 6 gives all three rules. Over the four draws the readout with a bias and a squared target has LDA error 0.088 to 0.137 on the four-class task, against 0.217 for the generic statistics, and 0.079 to 0.093 on the oscillators, against 0.129. The bias alone brings the bilinear family to 0.000 on every draw. On the AR against AR-GARCH task the plain readout, the readout with a bias alone and the AR coefficients are at chance, and with the squared target added the error is 0.117 to 0.206.

We next consider the squared target without a bias. Without a bias the state is odd in the input path, so for two classes symmetric under reflection the cross-moment of the state with $x_{t+1}^2 - 1$ has population value zero in both. For the AR and AR-GARCH pair, LDA on the coordinates of this target alone has error 0.50 to 0.52 at $n \in \{10, 20\}$ and $\lambda \in \{0.1, 1, 3\}$, and no coordinate has a class-mean difference larger than 0.12 pooled standard deviations. On the bilinear pair the same coordinates have error 0.010 to 0.061.

The AR and AR-GARCH classes also differ in higher-order moments of the state. We therefore also use a controlled pair of martingale-difference classes,

$$x_t = \varepsilon_t \sqrt{1 + a S_t}, \tag{F.7}$$

with iid Rademacher ε_t, $a = 0.75$, and S_t an independent stationary Markov chain on $\{-1, 1\}$ with $\mathbb{E}[S_{t+1} \mid S_t] = r S_t$, where $r = 0$ for one class and $r = 0.8$ for the other. Both classes have the marginal law with mass $\frac{1}{4}$ at each of $\pm\sqrt{1 \pm a}$, zero autocovariance at every non-zero lag and zero conditional mean, and they differ in

$$\mathbb{E}[(x_t^2 - 1)(x_{t+1}^2 - 1)] = a^2 r. \tag{F.8}$$

The population readout at the next-step target is zero for both, with or without a bias (Corollary 1), so no readout of the original target has a mean signal. The cross-moment of the state with the squared

target is $ar\,\mathbb{E}[S_t h_t]$, because $\mathbb{E}[x_{t+1}^2 - 1 \mid S_{\le t}, \varepsilon_{\le t}] = arS_t$. Without a bias it vanishes, since the state is odd in the signs ε. With a bias b and a memoryless unit with unit input weight,

$$\mathbb{E}[S_t h_t] = \tfrac{1}{2}\{g(1+a) - g(1-a)\}, \qquad g(v) = \sinh(2b)/(\cosh 2b + \cosh(2\lambda\sqrt{v})), \tag{F.9}$$

which is non-zero for $b \neq 0$ and $\lambda > 0$ because g is strictly monotone in v.

Table 5 gives two protocols at $L = 1024$ with 50 training sequences per class and four reservoir draws. The first uses the process as defined, which already has mean zero and unit variance, and fits an intercept in every variant, so that the bias and the squared target are the only differences between the variants. The plain readout, the bias alone and the squared target alone then have error between 0.48 and 0.52 at the selected scale and at every scale of the grid, and no readout coordinate has a class-mean difference above 0.23 pooled standard deviations. With both remedies the error is 0.000 and the class-mean difference is 6.7 to 15 standard deviations, the mean signal that the formula above gives. The second protocol standardizes every sequence and fits an intercept only with the bias, as for the families, and gives the same picture. The generic statistics, which contain $\overline{x_t^2 x_{t-1}^2}$, have error 0.000 in both.

Table 5: The controlled pair of martingale differences, $L = 1024$, 50 training sequences per class, scale chosen without labels, mean over four reservoir draws. Top: the process as defined, every variant with an intercept. "Error" is the shrinkage-LDA test error, "mean gap" the largest class-mean difference of a readout coordinate at $\lambda = 1$ in pooled standard deviations, range over draws. Bottom: every sequence standardized, intercept only with the bias, as in Table 6.

Protocol	n		plain	bias	x^2 target	bias, x^2 target
process, intercept	10	error	0.502	0.500	0.493	0.000
		mean gap	0.07–0.12	0.09–0.15	0.07–0.15	7.67–15.03
	20	error	0.492	0.495	0.488	0.000
		mean gap	0.13–0.23	0.11–0.16	0.13–0.23	6.74–10.90
standardized	10	error	0.491	0.499	0.490	0.000
	20	error	0.478	0.499	0.485	0.000

Table 6: Families of dynamical laws at $L = 1024$ with 50 training sequences per class, test error, mean over four reservoir draws.

Classes	rule	raw	generic	AR	readout	+bias	+bias, x^2 target
Four families	1-NN	0.732	0.275	0.625	0.665	0.357	0.220
	SVM	0.749	0.199	0.473	0.564	0.275	0.120
	LDA	0.749	0.217	0.458	0.500	0.277	0.105
Kuramoto, two channels	1-NN	0.326	0.162	0.271	0.103	0.104	0.108
	SVM	0.355	0.138	0.215	0.080	0.081	0.084
	LDA	0.548	0.129	0.208	0.082	0.082	0.085
AR against bilinear	1-NN	0.485	0.019	0.404	0.438	0.002	0.003
	SVM	0.519	0.006	0.300	0.362	0.000	0.000
	LDA	0.517	0.006	0.293	0.319	0.000	0.000
AR against AR-GARCH	1-NN	0.512	0.216	0.507	0.492	0.474	0.309
	SVM	0.514	0.112	0.508	0.488	0.396	0.162
	LDA	0.519	0.133	0.509	0.503	0.495	0.143
AR against threshold AR	1-NN	0.479	0.085	0.315	0.383	0.111	0.091
	SVM	0.490	0.035	0.136	0.242	0.041	0.035
	LDA	0.488	0.013	0.106	0.177	0.036	0.027

F.4 Entry orders

With the lower orders subtracted, the measured entry orders are within 0.017 of 3 and 5 (§5.1).

We verify the cubic ($k = 1$) and quintic ($k = 2$) instances of Theorem 3 at the self-step target by projecting the sample cross-moment gap onto $(W_{\mathrm{in}})^{\odot 3}$ and $(W_{\mathrm{in}})^{\odot 5}$, each Gram–Schmidt-orthogonalized against the lower odd powers of W_{in}. The raw gap is dominated by the fluctuation

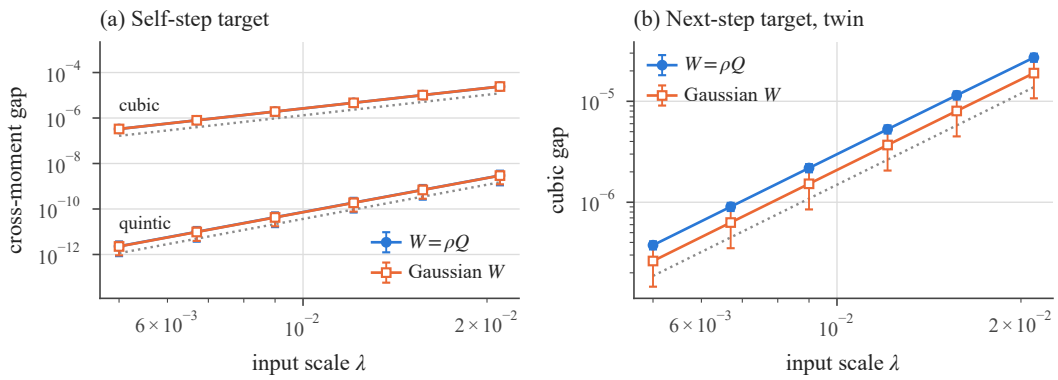


Figure 5: Entry orders (§5.1, Appendix F.4). *(a)* Control-variate projections of the self-step cross-moment gap on the cubic and quintic directions against λ, both arms, against slopes 3 and 5 (dotted). *(b)* The next-step cubic order on the twin, against slope 3. Error bars: one standard deviation over seeds. The dotted references are offset below the data.

of its lower orders. At $\lambda = 0.005$ its quintic projection is 5.0×10^{-7}, against a true value of 1.2×10^{-12}. We therefore subtract $\lambda\,\Delta\langle x_t a_t\rangle$ on both pairs and $\lambda^3\Delta\langle x_t c_t\rangle$ on the cum_6 pair, where a_t and $c_t = Wc_{t-1} - a_t^{\odot 3}/3$ are the linearized and cubic states of §E.1, computed on the same path. Both subtracted terms have population mean zero. The first vanishes because the three test laws share one autocovariance (Proposition C.1(a)). The second vanishes because on the cum_6 pair it depends only on moments of order at most four, which are matched. The estimator is therefore unbiased for the $\tanh$ gap, and its signal and noise share the scale λ^{2k+1}.

The estimator is applied to three iid test laws with zero mean and unit variance. The first is the standard Gaussian ($\mathrm{cum}_4 = \mathrm{cum}_6 = 0$), and the second is the standardized Laplace ($\mathrm{cum}_4 = 3$). The third is a symmetric four-point law that takes $\pm a$ with probability 0.4 each and $\pm b$ with probability 0.1 each, where $a^2 = 1 - 1/\sqrt{2}$ and $b^2 = 5 - 4a^2$. Then $\mu_2 = 1$ and $\mu_4 = 3$ (hence $\mathrm{cum}_4 = 0$, matched to the Gaussian), and $\mathrm{cum}_6 = -3.76$. The cubic pair is Gaussian against Laplace. The quintic pair is Gaussian against the four-point law, and these two laws are matched through fourth order.

The measurements use $n = 100$, $\rho = 0.9$, $T = 1000$, washout 100, $N_{\mathrm{seq}} = 8000$ per class and 6 seeds, on both reservoir arms of Figure 5(a). The three test laws are coupled by common random numbers with antithetic pairing. There is no per-sequence standardization. W_{in} is drawn Gaussian and rescaled to $\|W_{\mathrm{in}}\|_2 = 5.0$ exactly, so that every seed shares one drive grid. The grid $\lambda \in \{0.005, \ldots, 0.021\}$ keeps every drive $\lambda\|W_{\mathrm{in}}\|_2$ in $[0.025, 0.105]$. A λ point enters a slope fit only if the seed-averaged projected gap exceeds three times the seed-averaged projected sampling noise of the same estimator. Every point meets this criterion for both orders on both arms.

Result. With 95% bootstrap intervals over seeds, the fitted cubic slope is 2.994 $[2.993, 2.995]$ on both arms, against the asymptotic exponent 3. The quintic slope is 4.998 $[4.995, 5.008]$ on the dense Gaussian arm and 4.993 $[4.992, 4.996]$ on the orthogonal arm, against 5. The orthogonal arm satisfies Assumption 2 with $q = \rho$, and the dense arm gives the same exponents. The compensated ratio gap/λ^5 is constant to 0.3% and 1.1% across the grid on the two arms, while gap/λ^3 falls as λ^2. This is the empirical counterpart of the statement that cum_6 enters at λ^5 and no earlier.

We apply the same estimator to the next-step instance of Theorem 3 with $k = 1$, at span $s = 1$, with the twin of §F.1 as witness. Both classes filter one common-random-number Laplace stream, through the forward and the reversed coefficient vector respectively (the coefficients are divided by the population standard deviation, which reversal preserves). The pair therefore shares one autocovariance by construction, and the subtracted $\lambda\,\Delta\langle x_{t+1} a_t\rangle$ is zero in population by Proposition C.1(a). The projection direction is the coefficient itself. The λ-independent statistic $\Delta\langle x_{t+1} c_t\rangle$ is the empirical cubic coefficient, and its component along the fixed $(W_{\mathrm{in}})^{\odot 3}$ depends on the draw. We therefore project the gap on one half of the sequences onto the direction of $\Delta\langle x_{t+1} c_t\rangle$ estimated on the held-out other half. This keeps the direction independent of the noise it averages.

Result. The configuration is as above, with $N_{\mathrm{seq}} = 4000$ per class and 6 seeds. Every grid point meets the $3\times$ criterion, with signal-to-noise ratio above 70. Against the asymptotic exponent 3, the

fitted slopes are $2.987\ [2.984, 2.990]$ on the dense Gaussian arm and $2.983\ [2.983, 2.984]$ on the certified orthogonal arm. The ratio of the measured gap to the projected $\lambda^3\Delta\langle x_{t+1}c_t\rangle$ rises to 1 as λ falls, and it reaches 0.9989 and 0.9984 at $\lambda = 0.005$. This is the coefficient identity of Theorem 3(a) observed in the data. The ratio gap/λ^3 is constant to 2.0% and 2.6% across the grid on the two arms, which is the widest spread in this subsection.

F.5 Signal and noise of the cumulant channel

The measured covariance and separation are compared with Proposition D.2 (Section 4).

On the cumulant pair, iid Gaussian against Laplace at the self-step target, the constants in the proof of Proposition D.2(b) are explicit, although this pair has unbounded innovations. For iid classes that share their variance, the paired terms of $M^{(3)}$ agree and a joint fourth cumulant vanishes unless its four time indices coincide, so only the cube of the current input contributes to $\Delta M^{(3)}$, and the constants are

$$\Delta M^{(3)} = -\tfrac{1}{3}\Delta\mu_4\, W_{\mathrm{in}}^{\odot 3},$$
$$V_1 = (\mu_4 - \sigma^4) W_{\mathrm{in}} W_{\mathrm{in}}^\top + \sigma^4 \sum_{k\ge 1} W^k W_{\mathrm{in}} W_{\mathrm{in}}^\top (W^\top)^k. \tag{F.10}$$

The measurements use one draw at $n = 50$, with $W = 0.9Q$, unit-norm Gaussian W_{in}, washout 20 and 500 sequences per class. The grid is $\lambda \in \{0.005, 0.01, 0.02, 0.05, 0.1\}$, $T \in \{256, 1024, 4096, 16384\}$ and $\gamma \in \{10^{-2}, 10^{-1}, 1\}$. We run it on that pair and on the proposition's bounded pair, which is iid uniform against a symmetric four-point law taking $\pm\sqrt{13}/2$ with probability $\frac{1}{8}$ each and $\pm\frac{1}{2}$ otherwise ($\mu_4 = 1.8$ against 2.69). In each cell we measure three quantities.

The first is the separation along the Fisher direction $(\gamma I + \lambda^2\Sigma_{11})\bar{V}_1^{-1}\Delta M^{(3)\top}$, along which the separation equals d^2 of Proposition D.2(c), defined as the squared difference of the two sample means over the pooled sample variance. Its mean-estimation floor is $N_1^{-1} + N_2^{-1} = 0.004$. The second is the relative Frobenius error of the pooled covariance against $\lambda^2 R_\lambda \bar{V}_1 R_\lambda$. The third is the accuracy of LDA and of the nearest-centroid rule on a $50/50$ split.

Result. At $\gamma = 1$, where $\lambda^2\|\Sigma_{11}\|_{\mathrm{op}}/\gamma$ is about 0.01, the measured covariance is within 6.5–10.6% of $\lambda^2 R_\lambda \bar{V}_1 R_\lambda$ in every cell. This is the finite-sample floor of a 50×50 covariance estimated from 1000 sequences. At $\gamma = 10^{-2}$ the error reaches 40% and 76% at $\lambda = 0.05$ and 0.1, where $\lambda^2\Sigma_{11}$ is no longer small compared with γ. On the bounded pair, the separation in the five cells at $\lambda \le 0.05$ that exceed ten times the estimation floor is 0.91 to 1.51 times the leading order of Proposition D.2(c), and 0.42 to 0.51 times at $\lambda = 0.1$. The log-log slopes are 0.97 in the length, and 4.08 and 3.65 in the scale. LDA reaches accuracy 0.98 at $\lambda = 0.05$ and $T = 16384$, whereas the nearest-centroid rule stays at 0.47–0.51.

Under the relative ridge, on the cumulant pair with 300 sequences per class, $T \in \{1024, 4096\}$ and $g \in \{10^{-2}, 10^{-1}, 1\}$, the measured covariance of $\lambda\hat{W}$ is within 16% of the delta-method value of Remark D.4 at every (λ, g, T). In a separate run of 100 sequences of length 20000 per class, the analytic $\Delta M^{(3)} = -\frac{1}{3}\Delta\mu_4 W_{\mathrm{in}}^{\odot 3}$ agrees with the empirical mean gap of the cubic states to within 4%.

F.6 The constant of the separation law

The constant $\kappa_\star$ of the separation law is computed exactly, and the law is compared with measurement at finite scale (Figure 2).

We first compute the constant exactly. For the bounded iid pair of Appendix F.5 at the self-step target,

$$\Delta M^{(3)} = -\tfrac{1}{3}\Delta\mu_4 W_{\mathrm{in}}^{\odot 3},$$
$$\bar{V}_1 = (\bar{\mu}_4 - 2) W_{\mathrm{in}} W_{\mathrm{in}}^\top + \sum_{k\ge 0} W^k W_{\mathrm{in}} W_{\mathrm{in}}^\top (W^\top)^k. \tag{F.11}$$

For the reversal pairs at the next-step target we compute $\bar{V}_1$ from Bartlett's formula (D.18) and $\Delta M^{(3)}$ from the fourth cumulant of the innovations. The condition number of $\bar{V}_1$ is of order $10^2\rho^{-2n}$, so we compute in arithmetic with 50 to 80 digits.

For $n \geq 100$ we use the eigenbasis of Q, in which $\bar{V}_1$ is a Cauchy-like matrix determined by the eigenvalues of Q and the coordinates of W_{in}. The two routes agree to thirteen digits, and at $n = 5$ the next-step constants agree with a Monte Carlo estimate from 2×10^4 sequences per class to 1%. Table 7 gives the range over three to five draws of (Q, W_{in}) with a unit-norm Gaussian W_{in}.

Table 7: The constant $\kappa_\star = \Delta M^{(3)} \bar{V}_1^{-1} \Delta M^{(3)\top}$ of the bounded iid pair at the self-step target, $W = \rho Q$, range over draws. Dashes mark cells not computed.

ρ	$n = 10$	$n = 30$	$n = 50$	$n = 100$	$n = 200$
0.5	$1.5 \cdot 10^2$–$5.1 \cdot 10^3$	$7.4 \cdot 10^{11}$–$7.8 \cdot 10^{14}$	$8.6 \cdot 10^{24}$–$7.1 \cdot 10^{26}$	–	–
0.7	0.58–11	$6.6 \cdot 10^3$–$2.0 \cdot 10^6$	$7.3 \cdot 10^{10}$–$6.9 \cdot 10^{12}$	–	–
0.9	0.017–0.13	0.026–1.0	2.1–$2.9 \cdot 10^2$	$2.8 \cdot 10^4$–$5.5 \cdot 10^5$	$4.9 \cdot 10^{12}$–$7.2 \cdot 10^{13}$
0.95	0.0073–0.049	0.0023–0.058	0.011–1.6	0.71–21	$4.9 \cdot 10^3$–$1.6 \cdot 10^4$
0.99	0.0057–0.010	0.00035–0.0048	0.00059–0.0051	0.00032–0.017	0.00050–0.0013

On the two-tap pair with $W = \rho Q$, $\kappa_\star$ is between 0.006 and 0.05 at $n = 10$ and between 3.2 and 37 at $n = 50$ for $\rho = 0.9$, and between 27 and 1120 at $n = 10$ and between 4×10^{24} and 1.5×10^{26} at $n = 50$ for $\rho = 0.5$. On the five reservoir draws of the scale sweep (Appendix F.7, $n = 50$, $\rho = 0.9$) it is 2.5 to 12 on the two-tap pair, 3.3 to 93 on the eight-tap pair and 13 to 119 on the twin.

At finite scale, Table 8 compares the separation of the $\tanh$ readout with $\lambda^4 L \kappa_\star$ on the two-tap pair, for one draw at $n = 10$, $\rho = 0.5$ ($\kappa_\star = 1122$) and one at $n = 50$, $\rho = 0.9$ ($\kappa_\star = 5.6$), at the fixed ridge $\gamma = 1$ of Proposition D.2. The separation is the bias-corrected Mahalanobis estimate from 2000 sequences per class, and the errors are those of one training set against 1000 test sequences per class. Figure 2 shows the same two draws at $L \in \{256, 1024, 4096\}$.

At $n = 10$ the ratio is 1.02 to 1.10 from $\lambda = 0.01$ to 0.05 and falls to 0.87, 0.27, 0.018 and 0.0023 at $\lambda = 0.1$, 0.2, 0.5 and 1, while the separation per unit length keeps growing (0.098, 0.48, 1.23 and 2.55). This is the crossover of Figure 2a. At $n = 10$ and $L = 4096$, LDA trained on 1000 sequences per class has error 0.344 at $\lambda = 0.02$ and 0.002 at $\lambda = 0.05$, against the limit values 0.334 and 0.004 of Theorem 4(b). At smaller scales, $\lambda = 0.005$ at $n = 10$ and $\lambda \leq 0.02$ at $n = 50$, the separation is below the resolution of the estimate.

Table 8: Separation of the $\tanh$ readout at $L = 4096$ and $\gamma = 1$ against $\lambda^4 L \kappa_\star$, and test errors of LDA and of shrinkage LDA (LW), two-tap pair. Dashes: fewer training sequences than dimensions.

	λ	d^2	$\lambda^4 L \kappa_\star$	ratio	LDA, N=25	LDA, N=100	LW, N=25
$n = 10$, $\rho = 0.5$	0.02	0.810	0.735	1.10	0.424	0.354	0.492
	0.05	29.2	28.7	1.02	0.010	0.001	0.514
	0.1	400	459	0.87	0.000	0.000	0.499
	0.2	1974	7350	0.27	0.000	0.000	0.401
$n = 50$, $\rho = 0.9$	0.05	0.18	0.14	1.27	–	0.465	0.506
	0.1	2.02	2.30	0.88	–	0.354	0.497
	0.2	10.8	36.8	0.29	–	0.084	0.502
	0.5	46.5	1438	0.032	–	0.002	0.485

At $n = 50$, $\lambda = 0.1$ and $L = 4096$, LDA trained on 1000 sequences per class has error 0.251, against $\Phi(-\frac{1}{2}\sqrt{2.02}) = 0.238$ from the measured separation and the limit value $\Phi(-\frac{1}{2}\sqrt{2.30}) = 0.224$ of Theorem 4(b). Shrinkage LDA with 25 sequences per class stays between 0.48 and 0.53 at every scale up to $\lambda = 0.5$ at $n = 50$, and up to $\lambda = 0.1$ at $n = 10$, where LDA has error 0.000. Shrinkage toward the identity discards the directions of least linear noise, where the signal lies (Proposition D.5).

F.7 The plain readout across the input scale

Held-out prediction, without labels, selects input scales at which the plain readout separates the classes (Figure 3a).

Protocol. We use three pairs at the next-step target. Two are the two-tap and eight-tap pairs of Appendix F.1. The third is the twin of Appendix F.1 on standardized Laplace innovations, with its coefficients divided by the population standard deviation. The reservoirs have $n = 50$,

$W = \rho Q$ at $\rho = 0.9$ and 0.5, unit-norm Gaussian W_{in} and washout 20. The input scales are $\lambda \in \{0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100\}$, so the drive $\lambda\|W_{\text{in}}\|_2$ runs from 0.02 to 100. We use the fixed ridge $\gamma = 1$ and the per-sequence relative ridges $\gamma = g\,\mathrm{tr}(\hat{\Sigma}_h)/n$ at $g \in \{10^{-4}, 10^{-2}, 10^{-1}\}$. The lengths are $L \in \{256, 1024, 4096\}$ and the training sizes are $N \in \{25, 100\}$, drawn from a pool of 200 sequences per class, with 100 test sequences per class. Each cell uses five training subsets, and there are five independent draws of reservoir and data.

Held-out one-step R^2 is computed inside each sequence as in Appendix F.3. The selection statistic is its pooled value over the training sequences of the subset, namely one minus the summed residual over the summed variance. It does not use labels. The selected configuration is the (λ, g) with the largest mean, and at each λ the relative ridge is chosen in the same way. The classification rule is shrinkage LDA.

Results. Table 9 gives the results of shrinkage LDA at $N = 25$. At $\rho = 0.9$ the plain readout is at chance at $\lambda \leq 0.05$ in every cell. At $\lambda = 0.1$ it is also at chance in every cell except the eight-tap pair and the twin at $L = 4096$ (0.34 and 0.37). From $\lambda = 1$ it separates all three pairs. The error on the bounded pairs is 0.001 or less from $L = 1024$, and the error on the twin is 0.000 at $L = 4096$. At the saturated end the bounded pairs stay at or below 0.005 up to $\lambda = 100$. The twin, whose innovations are continuous, loses the information that saturation discards, and its errors at $\lambda = 100$ are 0.404, 0.249 and 0.068 over the three lengths.

On the two-tap pair at $L = 256$, pooled held-out R^2 rises from 0.11 at $\lambda \leq 0.5$ to 0.47 at $\lambda \geq 20$ (0.19 to 0.50 at $L = 4096$). On the eight-tap pair the statistic has an interior maximum near $\lambda = 20$, and on the twin near $\lambda = 1$ to 2. At $\rho = 0.9$ the selection chooses $\lambda \geq 0.5$ in 450 of 450 selections. The chosen scales are $\lambda = 20$ to 100 on the two-tap pair, 2 to 50 on the eight-tap pair and 0.5 to 5 on the twin. At $\rho = 0.5$ it does so in 433 of 450 selections.

At the selected configuration, across draws at $L = 256$ and $\rho = 0.9$ the errors range from 0.000 to 0.007 on the bounded pairs and from 0.156 to 0.266 on the twin. At $\rho = 0.5$ the error is 0.000 on both bounded pairs, and 0.135, 0.018 and 0.000 on the twin. At $\rho = 0.9$ on the twin, the error at the selected scale exceeds that at the best scale of the grid by 0.05 at $L = 256$ and 1024 (0.170 and 0.027 at $\lambda = 5$), and it is within 0.001 of it at $L = 4096$. Under shrinkage LDA, over all pairs, lengths and both values of N, the error at the scale selected by held-out R^2 exceeds the error at the best scale of the grid by at most 0.054 at $\rho = 0.9$ (twin, $L = 256$, $N = 25$) and 0.027 at $\rho = 0.5$.

Table 9: The plain readout across the input scale: shrinkage-LDA test error at $N = 25$, mean over five draws, next-step target. At each input scale the relative ridge is the one with the highest mean held-out R^2 over the training sequences. The column "sel." uses the (λ, g) that this statistic selects over the whole grid, without labels.

		$\rho = 0.9$ plain readout at $\lambda =$								$\rho = 0.5$ plain readout at $\lambda =$							
	L	0.02	0.1	0.5	1	2	10	50	sel.	0.02	0.1	0.5	1	2	10	50	sel.
two-tap	256	.495	.491	.339	.053	.000	.000	.000	.000	.507	.472	.000	.000	.000	.000	.000	.000
two-tap	1024	.477	.476	.137	.001	.000	.000	.000	.000	.495	.376	.000	.000	.000	.000	.000	.000
two-tap	4096	.518	.510	.014	.000	.000	.000	.000	.000	.499	.070	.000	.000	.000	.000	.000	.000
eight-tap	256	.499	.498	.365	.127	.021	.000	.001	.003	.472	.032	.003	.000	.000	.000	.002	.000
eight-tap	1024	.504	.472	.031	.001	.000	.000	.000	.000	.416	.000	.000	.000	.000	.000	.000	.000
eight-tap	4096	.512	.342	.000	.000	.000	.000	.000	.000	.255	.000	.000	.000	.000	.000	.000	.000
twin	256	.484	.485	.356	.235	.177	.203	.338	.223	.494	.402	.152	.124	.110	.145	.324	.135
twin	1024	.516	.512	.178	.066	.030	.039	.140	.076	.416	.071	.012	.014	.006	.027	.139	.018
twin	4096	.507	.370	.004	.000	.000	.000	.041	.001	.149	.003	.000	.000	.000	.000	.005	.000

F.8 THE SCALE AXIS WITH FINITE MEMORY

The population rule of finite-depth reservoirs is computed exactly and compared with the pipeline (Figure 3b).

Protocol. The reservoirs have $n = 24$ units in $p + 1$ layers of equal size, with $p \in \{0, 1, 2\}$. The matrix W has Gaussian entries from each layer to the next and zeros elsewhere, and is rescaled to $\|W\|_2 = 0.9$. The input vector W_{in} is Gaussian with unit norm, $\gamma = 1$, the target is one step ahead, and the washout is 10 steps. Besides the two-tap pair we use a pair whose difference lies at lag two, $x_t = (e_t + \frac{1}{2}e_{t-2})/\sqrt{5/4}$ against $(\frac{1}{2}e_t + e_{t-2})/\sqrt{5/4}$ with $e_t = \pm 1$. Its two classes have the same

law of (x_{t+1}, x_t), since x_{t+1} and x_t share no innovation. By Proposition D.6(a) the memoryless reservoir therefore has one population readout for both classes at every scale.

The computation enumerates the innovation windows of Proposition D.6(c). At each scale it gives the population readouts, the long-run covariance of the readout, the separation per unit length under a pooled covariance shrunk by 10% toward a multiple of the identity, and the population one-step R^2. It also gives the error at length L of LDA with this shrunk metric and the population parameters, in the Gaussian limit with the exact means and covariances. The pipeline uses shrinkage LDA on 25 training sequences per class, 400 test sequences per class, and three reservoir draws.

Before using the computation we check it. Identity (D.33) holds on single sequences to a relative error of 10^{-14} at $p = 0, 1, 2$ and $\lambda \in \{0.05, 1, 20\}$, with $(\hat{\Lambda}, \hat{Y})$ computed without the reservoir. Against 3000 simulated sequences per class of length 1000 at $p = 0, 1$, the exact long-run covariance agrees to within 4% at $\lambda = 0.3$ and 3. Where the mean gap exceeds its sampling error, at $\lambda = 3$, the exact gap agrees to within 1.2%. Because $\tanh$ is odd and the innovations are symmetric, $\mathfrak{h}_\lambda(-\varpi) = -\mathfrak{h}_\lambda(\varpi)$, and G_λ has rank at most half the number of window values. The pseudo-inverse form of (D.33) is therefore the one that applies here.

Results. Table 10 gives both errors at $L = 256$. At $L = 1024$ the pipeline error on the lag-two pair is 0.47 to 0.48 at every scale without memory, and 0.000 from $\lambda = 2$ with memory one or two. Without memory the population rule on the lag-two pair is at chance. The pipeline error there lies slightly below $\frac{1}{2}$ at $\lambda \geq 2$, because the long-run covariances of the readout differ between the classes, by about half their norm at $\lambda = 5$, and a linear rule with a shifted threshold can use this difference. On the two-tap pair at $\lambda \leq 0.5$ the pipeline error exceeds the error of the population rule, by up to 0.26 at $L = 1024$, because the metric of a 24-dimensional readout is estimated from 25 sequences per class.

Table 10: Pipeline error (shrinkage LDA, $N = 25$) and population LDA error in the Gaussian limit from the exact moments, on finite-depth reservoirs at $L = 256$, mean over three draws.

	lag-two pair, $p = 0$		lag-two pair, $p = 1$		lag-two pair, $p = 2$		two-tap pair, $p = 1$	
λ	pipeline	population	pipeline	population	pipeline	population	pipeline	population
0.2	0.503	0.500	0.491	0.494	0.480	0.477	0.490	0.333
0.5	0.504	0.500	0.401	0.454	0.361	0.330	0.139	0.017
1	0.499	0.500	0.287	0.279	0.098	0.033	0.000	0.000
2	0.470	0.500	0.052	0.064	0.000	0.000	0.000	0.000
5	0.476	0.500	0.000	0.000	0.000	0.000	0.000	0.000
20	0.476	0.500	0.000	0.000	0.000	0.000	0.000	0.000

F.9 Explicit higher-order statistics

Explicit statistics do better than the readout when they cover the lag and the order of the difference, and they are at chance beyond them (§5.1). We use the twenty reversibility statistics

$$\frac{1}{L}\sum_t (x_t^3 x_{t-k} - x_t x_{t-k}^3), \qquad k = 1, \ldots, 20, \tag{F.12}$$

a fourth-order version of the tests of Ramsey & Rothman (1996), which were built for time reversal.

To see where each family works as the lag of the difference grows, we use pairs whose difference lies at one lag J, namely $x_t \propto e_t + \frac{1}{2} e_{t-J}$ against its time reversal, with $e_t = \pm 1$ and $L = 1024$. We compare the twenty reversibility statistics with $k = 1, \ldots, 20$ and the readout of a reservoir of the same dimension, $n = 20$, $W = \rho Q$ with $\rho \in \{0.9, 0.95\}$, a unit-norm W_{in}, $\lambda = 1$ and the next-step target, with shrinkage LDA on 25 and LDA on 100 training sequences per class, five training sets and 500 test sequences per class (Table 11).

The reversibility statistics have error 0.000 for $J \leq 16$ and are at chance for $J \geq 24$, beyond their lags. The readout has error 0.031 or less for $J \leq 4$ at $\rho = 0.9$, rises to 0.24 to 0.29 at $J = 8$ and 0.38 to 0.40 at $J = 16$ over both reservoirs and rules, and is at chance for $J \geq 24$, in line with the lag envelope of Proposition E.3. On these pairs an explicit family that covers the lag of the difference does better than the readout, and both are at chance on a difference beyond their reach.

A family is also at chance when the order of the difference is not the one it was designed for. We give the two-tap pair innovations with the symmetric four-point law that takes $\pm\sqrt{11/3}$ with probability

Table 11: Test error on the single-lag pairs, $L = 1024$: reversibility statistics ($k \le 20$) and the readout ($n = 20$), each of dimension 20. LW: shrinkage LDA with 25 training sequences per class. LDA: 100 per class.

	reversibility		readout, $\rho = 0.9$		readout, $\rho = 0.95$	
J	LW	LDA	LW	LDA	LW	LDA
1	0.000	0.000	0.000	0.000	0.015	0.015
2	0.000	0.000	0.031	0.031	0.079	0.063
4	0.000	0.000	0.031	0.025	0.065	0.059
8	0.000	0.000	0.264	0.237	0.286	0.253
16	0.000	0.000	0.396	0.396	0.388	0.382
24	0.510	0.501	0.503	0.498	0.493	0.487
32	0.492	0.491	0.493	0.504	0.498	0.501
48	0.500	0.487	0.506	0.503	0.520	0.501

$\frac{9}{82}$ each and $\pm\frac{1}{2}$ with probability $\frac{16}{41}$ each, of unit variance and fourth moment 3, so that every joint fourth cumulant vanishes and the classes differ first at sixth order. At $L = 4096$ the twenty fourth-order reversibility statistics have error 0.491 with LDA on 100 training sequences per class, the same statistics with x^5 in place of x^3 have error 0.000, and the readout ($n = 20$, $\rho = 0.9$, next-step target) has error 0.481, 0.254, 0.021 and 0.000 at $\lambda = 0.5$, 1, 2 and 4. The difference enters the readout at order λ^5 (Theorem 3), so the scale decides whether it separates the pair.

F.10 The detection scale at order one

The order of a difference and the size of its witness set the scale at which it becomes detectable. We compute this scale exactly on layered reservoirs of memory $p \in \{1, 2\}$ with $n = 24$ units and $\|W\|_2 = 0.9$, a unit-norm Gaussian W_{in}, the next-step target and the fixed ridge $\gamma = 1$, for three draws each.

The innovations take finitely many values, so the population readouts and the long-run covariance of the fitted readout follow from Proposition D.6(c) by enumerating the innovation windows with their probabilities. Finite windows make this covariance singular in some directions. We therefore use the population form of shrinkage LDA: its direction is $a = \bar{V}_\eta^{-1}\Delta\bar{W}$, with $\bar{V}_\eta$ the pooled covariance shrunk toward its trace with weight $\eta = 0.1$, and its threshold is the midpoint of the class means. Its error at length L in the Gaussian limit is computed in the true class covariances V_1, V_2,

$$\tfrac{1}{2}\Phi\Big(-\frac{\sqrt{L}\,a^\top\Delta\bar{W}}{2\sqrt{a^\top V_1 a}}\Big) + \tfrac{1}{2}\Phi\Big(-\frac{\sqrt{L}\,a^\top\Delta\bar{W}}{2\sqrt{a^\top V_2 a}}\Big), \tag{F.13}$$

and its separation per unit length is $(a^\top\Delta\bar{W})^2/(a^\top\bar{V}a)$, with $\bar{V}$ the pooled covariance.

The pairs are the two-tap filter $(1, \vartheta_1)$ against its reversal, with $\vartheta_1 \in \{\frac{1}{2}, 0.8\}$ and four innovation laws of unit variance: ± 1 ($\mathrm{cum}_4 = -2$), the three-point law on $\{-\sqrt{2}, 0, \sqrt{2}\}$ with probabilities $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ ($\mathrm{cum}_4 = -1$), the Gauss–Hermite law of Appendix D.6 ($\mathrm{cum}_4 = 0$, $\mathrm{cum}_6 = -6$) and the four-point law of Appendix F.9 ($\mathrm{cum}_4 = 0$, $\mathrm{cum}_6 = -\frac{25}{6}$). The witness is

$$\begin{aligned} c_1(1) &= \mathrm{cum}_4(\vartheta_1 - \vartheta_1^3)/(1+\vartheta_1^2)^2 && \text{at fourth order,} \\ c_2(1) &= \mathrm{cum}_6(\vartheta_1 - \vartheta_1^5)/(1+\vartheta_1^2)^3 && \text{at sixth order.} \end{aligned} \tag{F.14}$$

On 25 scales from $\lambda = 0.05$ to 50 we fit the leading term $b\lambda^{4k}$ to this separation below $\lambda = 0.12$ and extrapolate it to $4\Phi^{-1}(0.9)^2/L$, the separation at which the rule has error 0.1 at $L = 1024$ in a shared covariance. The exact scale $\lambda_\star$ is where the error above falls to 0.1.

Table 12 gives the results at memory two. At memory one the ranges are similar: the slopes are 4.04 to 4.07 and 8.00 to 8.05, and the exact scale is 0.17 to 0.81 at fourth order and 0.74 to 4.03 at sixth order. Over both memories the exact scale is 0.78 to 3.34 times the extrapolation at fourth order, and 1.47 to 4.26 times at sixth order, where the extrapolation always comes too early. The Spearman correlation between the two over the 48 cases is 0.88, and within an order it is 0.91 at fourth and 0.73 at sixth order, where the witness alone gives 0.78 and 0.90.

A pair with a small fourth-order witness can need the scale of a sixth-order pair, as the three-point law at $\vartheta_1 = 0.8$ shows, and the leading coefficient, which scales with the witness, accounts for this.

The ratio of the exact scale to the extrapolation grows with the exact scale, with correlation 0.91 between their logarithms, so the extrapolation is more accurate when the detection scale is small. With shrinkage 0.1 the ratio is 0.71 to 5.95 at $L = 256$ and 0.83 to 2.75 at $L = 4096$, and with shrinkage 0.01 at $L = 1024$ it is 0.98 to 5.12.

Table 12: Exact detection scale on reservoirs of memory two ($n = 24$, fixed ridge $\gamma = 1$), range over three reservoirs. Slope: log-log slope of the separation below $\lambda = 0.12$. $\lambda_\star$: scale at which shrinkage LDA with the population parameters has Gaussian error 0.1 at $L = 1024$. λ_{lead}: the same scale from the extrapolated leading term. Last column: range of the ratio over the reservoirs.

Innovations	order	ϑ_1	$\lvert c_k(1)\rvert$	slope	$\lambda_\star$	λ_{lead}	$\lambda_\star/\lambda_{\mathrm{lead}}$
± 1	4	0.5	0.480	4.05–4.06	0.22–0.34	0.22–0.37	0.92–0.97
± 1	4	0.8	0.214	4.05–4.06	0.32–0.50	0.34–0.59	0.83–0.95
three-point, $\mathrm{cum}_4 = -1$	4	0.5	0.240	4.02–4.05	0.34–0.48	0.33–0.53	0.91–1.10
three-point, $\mathrm{cum}_4 = -1$	4	0.8	0.107	4.02–4.06	0.72–1.67	0.50–0.83	1.17–3.34
Gauss–Hermite	6	0.5	1.440	8.01–8.04	0.94–1.31	0.57–0.86	1.53–1.65
Gauss–Hermite	6	0.8	0.642	8.01–8.04	1.48–2.10	0.72–1.07	1.96–2.06
four-point	6	0.5	1.000	7.99–8.03	1.31–1.87	0.63–0.94	1.99–2.09
four-point	6	0.8	0.446	7.99–8.03	3.15–4.39	0.79–1.18	3.73–4.26

A simulation on a reservoir with full memory, which Remark D.7 relates to its finite-memory truncations, also places the onset of detection at order one. With $n = 20$, $W = 0.9Q$, a unit-norm Gaussian W_{in}, the fixed ridge $\gamma = 1$, $\vartheta_1 = \frac{1}{2}$, $L = 1024$, 1500 sequences per class and three reservoir draws, LDA with 100 training sequences per class reaches error 0.1 near $\lambda = 0.80$ for ± 1 innovations, 1.9 for the three-point law, 1.8 for the Gauss–Hermite law and 3.0 for the four-point law. With the relative ridge $g = 10^{-3}$ instead, the ± 1 and four-point pairs reach it near $\lambda = 0.67$ and 2.8. Unlike on the finite-memory reservoirs, where the three-point pair at $\vartheta_1 = \frac{1}{2}$ is separated before the Gauss–Hermite pair, here it comes after it.

F.11 Separation per unit length against explicit statistics

The factor 1.4 to 20 of §5.1 comes from this comparison of separation per unit length.

For the two-tap and eight-tap pairs with Rademacher innovations, $L = 1024$ and 2000 sequences per class, we compute the bias-corrected Mahalanobis separation of the class means, divided by L, for three representations: the next-step readout ($n = 20$, $W = \rho Q$, unit-norm Gaussian input weights, fixed ridge $\gamma = 10^{-3}$), the reversibility statistics

$$\frac{1}{L}\sum_t (x_t^3 x_{t-k} - x_t x_{t-k}^3) \quad \text{for } k = 1, \dots, 20, \tag{F.15}$$

and the 40 lag moments $\overline{x_t^3 x_{t-k}}$ and $\overline{x_t x_{t-k}^3}$ for $k = 1, \dots, 20$.

The two explicit families have the same separation per unit length to three digits, 38.4 on the two-tap pair and 0.440 on the eight-tap pair. Table 13 gives the ratio of the readout's separation to theirs. At $\rho = 0.5$ the ratio is largest at $\lambda = 1$ on the two-tap pair and at $\lambda = 3$ on the eight-tap pair for every draw, and it falls at larger scales, where the state saturates. At $\rho = 0.9$ it is below 0.002 up to $\lambda = 0.3$ and still grows at $\lambda = 30$ on the two-tap pair. With binary innovations the covariance of the readout becomes singular at $\lambda \geq 100$, where the separation is no longer defined.

Table 13: Separation per unit length of the readout ($n = 20$) relative to twenty fourth-order reversibility statistics, $L = 1024$, mean over three reservoir draws. Last column: the largest ratio over the grid, range over the draws.

Pair	ρ	$\lambda = 0.1$	0.3	1	3	10	30	best
Two-tap	0.5	0.062	0.120	**0.237**	0.106	0.060	0.073	0.14–0.35
	0.9	0.000	0.000	0.001	0.014	0.034	**0.079**	0.05–0.12
Eight-tap	0.5	0.173	0.258	0.388	**0.640**	0.482	0.469	0.54–0.70
	0.9	0.000	0.001	0.062	0.264	0.288	**0.323**	0.23–0.41