DRIFT IS NOT ENOUGH: A DIFFUSIVE FLOOR FOR GRADIENT PROPAGATION IN RECURRENT NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Whether a recurrent model learns long-range dependencies is determined by whether its gradients survive backpropagation over long horizons without exploding or vanishing. This behavior is governed by the top Lyapunov exponent λ_1, which characterizes the asymptotic rate of gradient growth or decay. While recent regularization methods seek to ensure marginal stability by driving this exponent to zero, the dynamics governing the non-autonomous gradient propagation remain insufficiently characterized. In this work, we model gradient dynamics as a linear cocycle over an ergodic Markov chain. Within this framework, we prove that the Lyapunov exponent spectrum of Recurrent Neural Network (RNN) gradient dynamics is well-defined and independent of the input realization, acting as a deterministic Law of Large Numbers drift. Crucially, while this drift rate governs the asymptotic scaling, gradient transport over any finite horizon m exhibits second-order fluctuations governed by a companion Central Limit Theorem with asymptotic variance σ^2. Consequently, eliminating drift ($\lambda_1 = 0$) does not prevent degradation, but leaves a superpolynomial, two-sided dispersion scaling as $e^{\pm\sigma\sqrt{m}}$, rendering vanishing and explosion equiprobable. We analytically prove that the asymptotic variance is strictly non-degenerate ($\sigma > 0$) for any choice of recurrent weights in vanilla RNNs with saturating activations, showing that gradient dispersion is an inescapable feature of these models. In our experiments, penalizing both the drift λ_1 (gradient flossing) and the variance σ^2 reduces empirical diffusion by 63–74% and improves psMNIST accuracy from 0.811 (unregularized) and 0.834 (flossing alone) to 0.860. Because this formulation requires only a recurrent state, it extends beyond vanilla RNNs, and we observe the same $\sigma\sqrt{m}$ growth in gated, orthogonal and selective state-space recurrent models.

1 INTRODUCTION

Whether a recurrent model learns long-range dependencies is determined by whether backpropagated gradients survive over long horizons without exploding or vanishing. Historically, this failure mode has been analyzed through exponential growth or decay rates, motivating methods that control the top Lyapunov exponent of the hidden dynamics,

$$\lambda_1 \;=\; \lim_{m\to\infty} \tfrac{1}{m} \log \left\| J_{t\to t+m} \right\|, \tag{1}$$

where $J_{t\to t+m} := \frac{\partial h_{t+m}}{\partial h_t}$ denotes the hidden-to-hidden transition Jacobian accumulated over m steps. A network with $\lambda_1 = 0$ operates at *marginal stability*, or the edge of chaos. A growing family of regularizers seeks to promote this operating point; for example, *gradient flossing* penalizes the squared magnitude of the top k Lyapunov exponents, $\sum_{i=1}^{k} \lambda_i^2$, driving them toward zero via differentiable QR iterations (Engelken, 2023), while related works examine Lyapunov spectra to diagnose trainability (Vogt et al., 2022; Engelken et al., 2023; Vogt et al., 2024). The benefits of operating near $\lambda_1 \approx 0$ have a rich history across reservoir computing and theoretical neuroscience (Bertschinger & Natschläger, 2004; Legenstein & Maass, 2007; Schuecker et al., 2018).

While setting $\lambda_1 \approx 0$ removes the average asymptotic exponential growth or decay, existing Lyapunov regularizers pursue this condition without a formal characterization of the underlying non-autonomous gradient dynamics. In particular, because the exponents are defined as asymptotic limits along specific

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.

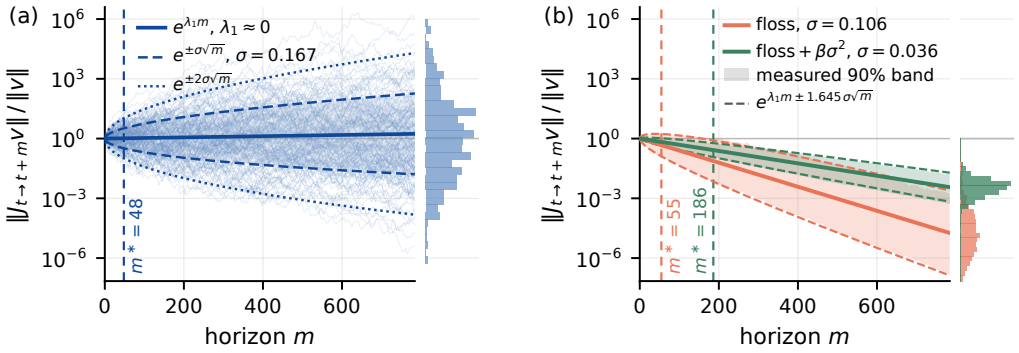

Figure 1: **Setting $\lambda_1 = 0$ removes the drift of the backpropagated signal, not its spread.** Gray: realizations of $\|J_{t\to t+m}v\|/\|v\|$; vertical strip: their distribution at $m = 784$ (512 trajectories). *(a)* Untrained vanilla RNN ($d_h = 128$, i.i.d. inputs) at $\lambda_1 = 0$. The mean log-magnitude is flat (black), but the trajectories follow the envelopes $e^{\pm z\sigma\sqrt{m}}$, $z \in \{1, 2\}$, of Theorem 4.1, and the central 90% spans 7 decades. *(b)* Trained networks differing only in the variance penalty: the spread falls from 4.1 decades (flossing) to 1.5 (with $\beta\sigma^2$), and the usable horizon m^* more than triples.

trajectories, their validity as deterministic optimization objectives requires establishing that they do not depend on the initial state or input realization. In this paper, we fill this gap by formulating gradient dynamics as a linear cocycle over an ergodic Markov shift. Within this framework, we prove that the entire Lyapunov spectrum is deterministic and well-defined almost surely, depending solely on the network parameters and the input distribution (cf. Theorem 3.1). The top exponent λ_1 thus acts as a deterministic Law of Large Numbers (LLN) drift rate.

Importantly, the Law of Large Numbers characterizes only the asymptotic growth rate. Over any finite horizon m, backpropagation along individual sample paths exhibits substantial deviations around this deterministic drift. Along a specific realization, finite-horizon gradient transport is governed by a companion Central Limit Theorem (cf. Theorem 4.1),

$$\log \frac{\left\|J_{t\to t+m}\, v\right\|}{\|v\|} = \underbrace{\lambda_1\, m}_{\text{drift}} + \underbrace{\sigma\sqrt{m}\, Z_m}_{\text{diffusion}}, \qquad Z_m \xrightarrow{d} \mathcal{N}(0,1), \tag{2}$$

where λ_1 represents the deterministic drift and σ characterizes the asymptotic diffusion scale. This reveals that setting the drift to zero ($\lambda_1 = 0$) does not prevent gradient degradation; instead, it leaves a two-sided, superpolynomial dispersion scaling as $e^{\pm\sigma\sqrt{m}}$. This insight rectifies a prevalent assumption in the literature that marginal stability yields polynomial or square-root gradient scaling (Kerg et al., 2019). The square root applies to the *logarithm* of the gradient norm, so the rescaling factor exceeds any polynomial with probability approaching one-half, and falls below any inverse polynomial with the same probability (Figure 1).

The gradient dispersion admits an irreducible lower bound (a *diffusive floor* $\sigma > 0$) that cannot be removed in vanilla RNNs. We prove analytically that for every vanilla RNN with a saturating activation and a non-degenerate input law, $\sigma > 0$ for any parameter choice, including orthogonal and unitary restrictions (Proposition 4.2). Gradient propagation is therefore described by two parameters, the drift λ_1 and the diffusion σ. Together they determine a *usable horizon* m^*, the number of steps over which a transported gradient stays within a given tolerance with high confidence (Corollary 4.3); when the drift is small ($\lambda_1 \to 0$), m^* is limited by σ. The diffusion can be estimated and penalized through the same QR recursion used for λ_1, at negligible cost. In our experiments, $m^* \geq H_{\text{task}}$, where H_{task} is the horizon a task requires, behaves as a graded necessary condition that can be checked early in training, and reducing σ beyond what the task requires lowers accuracy.

Contributions. The contributions of the paper are as follows:

(i) **Ergodic Cocycle Formulation of Recurrent Gradients.** We model the gradient dynamics of input-driven recurrent networks as a linear cocycle over a uniquely ergodic Markov shift, establishing a uniform Doeblin minorization that yields a unique ergodic stationary driving measure for non-autonomous gradient propagation.

(ii) **Deterministic Spectrum and Central Limit Theorem for RNNs.** We prove that the Lyapunov spectrum is a deterministic function of network parameters and input law (Theorem 3.1), and that gradient transport obeys a Central Limit Theorem with asymptotic variance σ^2 (Theorem 4.1).

(iii) **Strictly Non-Degenerate Diffusive Floor.** We prove that $\sigma > 0$ for any vanilla RNN with a saturating activation, regardless of W_h (Proposition 4.2), so that at $\lambda_1 = 0$ superpolynomial explosion and vanishing are equally likely, and derive the usable horizon $m^*(\kappa, \alpha)$ determined by (λ_1, σ) (Corollary 4.3).

(iv) **Variance-Aware Regularization and the Usable Horizon.** We observe the $\sigma\sqrt{m}$ growth in every recurrent cell we test whose Jacobian depends on the input, not only in vanilla RNNs. We estimate σ^2 at a few percent of additional cost over regular flossing, and show that at matched drift, penalizing σ^2 improves training on psMNIST and the adding problem, and controlling λ_1 alone is not sufficient. The usable horizon, compared with the horizon a task requires, behaves as a necessary condition for learning, and reducing σ beyond what the task requires lowers accuracy.

1.1 Related work

Lyapunov exponents as a training signal. Engelken (2023); Engelken et al. (2023) investigated the Lyapunov spectra of recurrent networks and introduced gradient flossing to align spectra near criticality, building on differentiable QR estimators proposed by Vogt et al. (2022; 2024). These formulations assume the deterministic nature of the spectrum that we formalize in Section 3. Mikhaeil et al. (2022) analyzed loss gradients via the spectra of RNN-generated orbits, but focused on autonomous attractors rather than input-driven systems. While Avelin & Karlsson (2022) and Borkar (2026) applied multiplicative ergodic theory to feedforward depth (where Jacobians become i.i.d. asymptotically), recurrent architectures maintain temporal dependencies across states, requiring the Markov-shift cocycle formulation developed in Section 3. Classical limit theorems for products of independent random matrices (Le Page, 1982; Bougerol & Lacroix, 1985) do not extend to the state-dependent products of recurrent models, motivating our construction of the projective chain in Section 4.

Architectures that constrain the spectrum. Pointwise matrix constraints have been extensively explored to control gradient norms, including unitary (Arjovsky et al., 2016; Jing et al., 2017), orthogonal (Henaff et al., 2016; Vorontsov et al., 2017; Helfrich et al., 2018), antisymmetric (Chang et al., 2019), Lipschitz (Erichson et al., 2021), non-normal (Kerg et al., 2019; Orhan & Pitkow, 2020), and coupled-oscillator (Rusch & Mishra, 2021) parameterizations, alongside dynamical isometry (Pennington et al., 2017; Chen et al., 2018; Hanin & Nica, 2020). These pointwise bounds limit each one-step stretch but not its fluctuation, and do not make σ zero (Appendix D.6). For selective state-space (SSMs) (Gu et al., 2022; Gu & Dao, 2023), we discuss the relation between λ_1 and σ in Section 6. In input-driven systems, ergodicity has been studied via the echo-state property (Manjunath & Jaeger, 2013; Wainrib & Galtier, 2016; Ortega & Rossmannek, 2025) and nonlinear autoregression (Trapletti et al., 2000); our contribution bridges these state dynamics to the asymptotic stability of the linear gradient cocycle. Appendix A provides an extended discussion of this line of work.

2 Setup

We analyze gradient propagation through the product of one-step Jacobians accumulated along an input stream. While this formulation applies to any model maintaining a recurrent hidden state, we focus on the prototypical vanilla recurrent neural network with hidden dimension d_h, input dimension d_u, and initial state $h_0 \in \mathbb{R}^{d_h}$. Driven by an input sequence $\{u_t\}_{t\geq 1} \subset \mathbb{R}^{d_u}$, the hidden state evolves according to

$$h_t \;=\; \phi(a_t), \qquad a_t \;:=\; W_h h_{t-1} + W_u u_t + b_h, \qquad t \geq 1, \tag{3}$$

where $W_h \in \mathbb{R}^{d_h \times d_h}$, $W_u \in \mathbb{R}^{d_h \times d_u}$, $b_h \in \mathbb{R}^{d_h}$, and the activation function ϕ acts coordinate-wise. Differentiating equation 3 yields the one-step Jacobian $J_t := \partial h_t / \partial h_{t-1} = D_t W_h$, where $D_t := \mathrm{diag}(\phi'(a_t)) \in \mathbb{R}^{d_h \times d_h}$ denotes the diagonal matrix with vector $\phi'(a_t)$ along its principal diagonal. Over a horizon of m steps, the long-term Jacobian is

$$J_{t\to t+m} \;:=\; \frac{\partial h_{t+m}}{\partial h_t} \;=\; J_{t+m} J_{t+m-1} \cdots J_{t+1}. \tag{4}$$

Backpropagation through time transports error covectors via the transpose $\partial\mathcal{L}/\partial h_t = J_{t\to T}^{\top}\,\partial\mathcal{L}/\partial h_T$ for any scalar loss $\mathcal{L}$ evaluated at the terminal state h_T. Because matrix transposition preserves the singular spectrum, the asymptotic scaling of the backward pass is governed by the ordered singular values $\sigma_1(J_{t\to t+m}) \geq \cdots \geq \sigma_{d_h}(J_{t\to t+m})$. We characterize this scaling through the finite-time Lyapunov exponents

$$\lambda_i^{(m)}(t) \;:=\; \tfrac{1}{m}\log\sigma_i\big(J_{t\to t+m}\big), \qquad i = 1,\dots,d_h, \tag{5}$$

where $\sigma_i(\cdot)$ denotes the i-th singular value operator (not to be confused with the scalar diffusion parameter σ established in Section 4). In particular, $\lambda_1^{(m)}(t) = \frac{1}{m}\log\|J_{t\to t+m}\|$, where $\|\cdot\|$ denotes the Euclidean vector norm and the induced matrix spectral 2-norm throughout. Gradient flossing penalizes $\sum_{i\leq k}(\lambda_i^{(m)}(t))^2$ during optimization.

Standing assumptions.

Assumption 1 (Inputs)**.** The inputs $\{u_t\}_{t\geq 1}$ are i.i.d. with law Γ on $\mathbb{R}^{d_u}$, absolutely continuous with density γ strictly positive everywhere and lower semicontinuous.

Assumption 2 (Activation)**.** $\phi : \mathbb{R} \to (-1,1)$ is a C^1 diffeomorphism with $\phi(0) = 0$ and $0 < \phi' \leq 1$.

Assumption 3 (Weights: ℓ-step controllability)**.** ϕ is C^∞, W_h is nonsingular, and for some $\ell \geq 1$: *(a)* for every $x \in \mathcal{X}$, there exists an input sequence $(u_1,\dots,u_\ell)$ generating pre-activations $a_1,\dots,a_\ell$ from $h_0 = x$ with $D_i := \mathrm{diag}(\phi'(a_i))$ such that

$$\mathrm{rank}\,\big[\; W_u \quad W_hD_{\ell-1}W_u \quad \cdots \quad W_hD_{\ell-1}W_h\cdots D_1W_u \;\big] \;=\; d_h; \tag{6}$$

and *(b)* an equilibrium $h^\star = \phi(W_hh^\star + W_uu^\star + b_h) \in (-1,1)^{d_h}$ is reachable from every $x \in \mathcal{X}$.

Assumption 2 holds for standard sigmoidal activations such as $\tanh$ and confines the hidden state to the compact cube $\mathcal{X} := [-1,1]^{d_h}$ after one step ($h_0 \in \mathcal{X}$). Unbounded activations (e.g., ReLU) preserve the cocycle identity but prevent the Doeblin minorization by failing to generate a positive transition density. The condition in equation 6 generalizes the classical Kalman controllability rank condition (Kalman, 1963) to non-autonomous trajectories. Because the diagonal matrices D_i reflect the realized saturation profile rather than static parameters, full-rank gradient propagation requires steering trajectories away from heavily saturated regions of the state space. The block matrix in equation 6 contains ℓd_u columns; thus, the rank condition can be satisfied for any input dimension $d_u \geq 1$ provided $\ell \geq \lceil d_h/d_u \rceil$ (cf. Remark C.4).

3 Well-posedness of the Lyapunov spectrum

When characterizing recurrent gradient cocycles, the Lyapunov exponents $\lim_{m\to\infty}\lambda_i^{(m)}(t)$ provided by the multiplicative ergodic theorem are a priori trajectory-dependent random variables—shift-invariant functions of the input realization rather than deterministic scalars (Furstenberg & Kesten, 1960; Oseledets, 1968). If the underlying driving dynamics lack ergodicity, these exponents fluctuate across sample paths and initial states, rendering Lyapunov-based regularizers ill-posed as optimization objectives. Proving that the spectrum collapses to deterministic constants depending solely on the network parameters and the input law is therefore critical. Establishing this well-posedness is non-trivial: the one-step Jacobians $J_t = D_tW_h$ cannot be analyzed as an i.i.d. matrix product because the activation derivatives D_t couple to the entire input history $\{u_s\}_{s\leq t}$. The following result addresses this challenge.

Theorem 3.1 (Deterministic Lyapunov spectrum of RNN gradients)**.** *Let Assumptions 1, 2, and 3 hold for the recurrent network equation 3. There exist deterministic constants $\lambda_1 \geq \cdots \geq \lambda_{d_h}$ in $[-\infty,\infty)$, uniquely determined by $(W_h, W_u, b_h, \phi, \Gamma)$, such that for every initial state $h_0 \in \mathcal{X}$ and every time origin $t \geq 0$, for almost all input sequences $\{u_s\}_{s\geq 1}$,*

$$\lim_{m\to\infty}\lambda_i^{(m)}(t) \;=\; \lambda_i, \qquad i = 1,\dots,d_h. \tag{7}$$

In particular, $\lim_{m\to\infty}\frac{1}{m}\log\|J_{t\to t+m}\| = \lambda_1$ almost surely, with $\lambda_1 \leq \log\|W_h\|$.

Theorem 3.1 ensures that sample-path estimators are strongly consistent and invariant to the initial state h_0, origin t and input sequence $\{u_s\}_{s\geq 0}$.

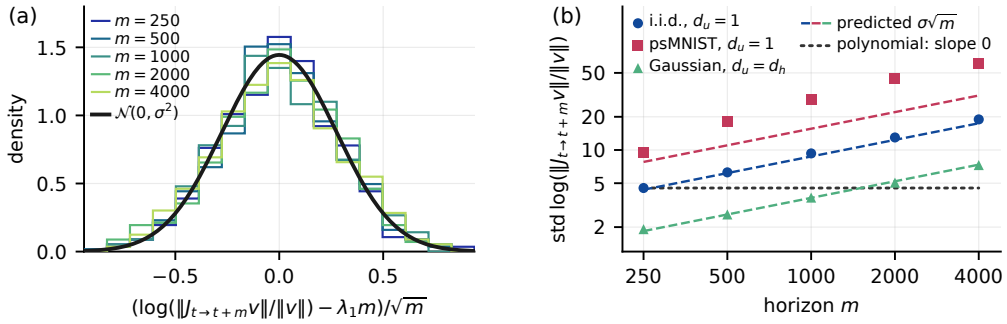

Figure 2: **Empirical validation of the Central Limit Theorem (Theorem 4.1). (a)** Empirical distribution of the centered, $\sqrt{m}$-scaled log-gradient ratio $(\log(\|J_{t\to t+m}v\|/\|v\|) - \lambda_1 m)/\sqrt{m}$ evaluated across horizons $m \in [250, 4000]$, compared against the predicted asymptotic normal density $\mathcal{N}(0, \sigma^2)$ (black curve). The scale σ is independently computed via the batch-means estimator without distribution fitting. **(b)** Dispersion width, the standard deviation of $\log(\|J_{t\to t+m}v\|/\|v\|)$ over 512 trajectories, versus horizon m (log–log scale). Under both 1D i.i.d. inputs ($d_u = 1$, blue circles) and full-rank Gaussian inputs satisfying all hypotheses of Theorem 4.1 ($d_u = d_h = 128$, green triangles, a network calibrated to $\lambda_1 \approx 0$), empirical scaling matches the predicted $\sigma\sqrt{m}$ rate (slopes 0.52 and 0.48 vs. theoretical 0.50, dashed lines). Under correlated sequence data (psMNIST, red squares), temporal dependence inflates dispersion (slope 0.66) while preserving diffusive scaling. All regimes reject the horizon-independent width that polynomial scaling would imply (slope 0, black dotted line).

Proof strategy. To recover Markovian structure, we enlarge the phase space to $\mathcal{Z} := \mathcal{X} \times \mathbb{R}^{d_u}$ with state $z_t := (h_{t-1}, u_t)$. The gradient product equation 4 is then identically the linear cocycle generated by

$$A(h, u) \;:=\; \mathrm{diag}\big(\phi'(W_h h + W_u u + b_h)\big) W_h \tag{8}$$

over the Markov chain $\{z_t\}_{t\geq 1}$ (Proposition C.1). Under Assumption 3, ℓ-step controllability ensures forward accessibility, making $\{h_t\}$ a T-chain, while the reachable equilibrium supplies ψ-irreducibility and aperiodicity. Compactness of $\mathcal{X} = [-1, 1]^{d_h}$ then makes the entire state space small, establishing a uniform ℓ-step Doeblin minorization with a unique invariant measure $\pi \otimes \Gamma$ and an ergodic path measure μ (Proposition C.2). Combining this base ergodicity with the multiplicative ergodic theorem yields constant exponents μ-almost surely; total variation convergence then transfers this determinism to every initial state h_0 and origin t (Lemma C.6).

4 Beyond Drift: The Diffusive Floor

While Theorem 3.1 establishes the deterministic Law of Large Numbers drift of gradient propagation, it characterizes only the first-order exponential rate. Backpropagation along individual sample paths is subject to trajectory-level fluctuations governed by a Central Limit Theorem (CLT). In this section, we establish this CLT, prove that its asymptotic variance is strictly positive across all vanilla RNN weight configurations, and derive the resulting usable gradient horizon. Complete proofs are deferred to Appendix D.

Assumption 4 (Full-rank input control)**.** W_h is nonsingular and $W_u \in \mathbb{R}^{d_h \times d_u}$ has full row rank $(\mathrm{rank}(W_u) = d_h \leq d_u)$.

Assumption 4 strengthens the multi-step condition of Assumption 3 to single-step ($\ell = 1$), ensuring that the conditional distribution of pre-activations $a_t \mid h_{t-1}$ admits a strictly positive, lower semi-continuous density across $\mathbb{R}^{d_h}$ (Lemma D.4) and that Theorem 3.1 holds directly via a one-step Doeblin condition (Lemma D.1). This input controllability guarantees geometric ergodicity for the joint evolution of hidden states and transported gradient directions. Although standard sequence tasks mostly have $d_u < d_h$, the empirical diffusion rate tests remain robust to this real-task condition, matching theoretical predictions even at $d_u = 1$ (Figure 2). Additional regularity conditions on activation profiles and input moments (Assumptions 5–7) are detailed in Appendix D; they hold generically for $\tanh$ networks driven by sub-Gaussian noise with dense W_h.

4.1 Central Limit Theorem for the Gradient Dynamics

To characterize trajectory-level fluctuations during backpropagation, we analyze the directional transport $\|J_{t\to t+m}v\|/\|v\|$ of an arbitrary error vector $v \neq 0$. The following result establishes that these sample-path fluctuations obey a Central Limit Theorem centered at the deterministic drift λ_1.

Theorem 4.1 (Central Limit Theorem for the Gradient Dynamics)**.** *Under Assumptions 1, 2, 4, and 5–7, $\lambda_1 \in (-\infty, \infty)$ and there exists a deterministic constant $\sigma \geq 0$, depending solely on $(W_h, W_u, b_h, \phi, \Gamma)$, such that for every initial state $h_0 \in \mathcal{X}$, time origin $t \geq 0$, and non-zero vector $v \in \mathbb{R}^{d_h}$,*

$$\frac{\log\left(\|J_{t\to t+m}\, v\| \,/\, \|v\|\right) - \lambda_1 m}{\sqrt{m}} \xrightarrow{d} \mathcal{N}(0, \sigma^2) \quad \text{as } m \to \infty. \tag{9}$$

Theorem 4.1 characterizes the trajectory-level dispersion of backpropagation: λ_1 dictates the asymptotic drift, while σ quantifies pathwise fluctuations. An RNN can satisfy $\lambda_1 = 0$ either through uniformly isometric transport or by averaging sub-exponential (yet superpolynomial) exploding and vanishing runs, depending on the input sequence. Standard drift-based Lyapunov regularizers cannot distinguish between these two regimes.

Proof strategy: Unlike the asymptotic drift λ_1—which follows from Kingman's subadditive ergodic theorem on the matrix cocycle—establishing a Central Limit Theorem requires an exact *additive* functional. Because the one-step stretching factor $\log(\|J_{t+n}v_{t+n-1}\|/\|v_{t+n-1}\|)$ depends on the orientation of the transported vector, we must track the joint evolution of the hidden state and the projective ray $[v_t] := \mathbb{R}v_t \in \mathbb{P}^{d_h-1}$ on the bundle $\mathcal{W} := \mathcal{X} \times \mathbb{P}^{d_h-1}$. The log-norm ratio then decomposes into the additive functional in equation 30. In classical products of random matrices, central limit theorems require the matrix semigroup to act strongly irreducibly and contractingly on projective space (Bougerol & Lacroix, 1985). In recurrent networks, contraction is supplied directly by activation saturation: when a pre-activation entry $|a_i|$ is moderate while $|a_j| \gg 1$ for all $j \neq i$, the diagonal matrix $D(a) = \mathrm{diag}(\phi'(a))$ concentrates along the coordinate axis e_i. Consequently, $D(a)W_h$ approaches rank one, collapsing all directions simultaneously onto $[e_i]$ (Lemma D.6). Alternating phases of collapse and spread establish a uniform Doeblin minorization on projective space (Proposition D.8).

Figure 2 verifies Theorem 4.1 empirically. Under i.i.d. inputs, standardized log-gradient norms match $\mathcal{N}(0, \sigma^2)$ across horizons $m \in [250, 4000]$, exhibiting $\sigma m^{0.52}$ width scaling. This refutes prevalent assumptions in the literature (e.g., Kerg et al., 2019) that marginal stability yields polynomial gradient scaling, which would predict a horizon-independent log-width (slope 0 in Figure 2b). When driven by temporally correlated inputs (e.g., psMNIST), persistent temporal structure broadens the dispersion (exponent ≈ 0.66), while maintaining its diffusive nature.

4.2 RNN Strict Positivity of the Diffusive Floor and Usable Horizon

The fundamental failure mode of marginal stability emerges because σ cannot be eliminated in RNN architectures.

Proposition 4.2 (Strict Non-Degeneracy of the Diffusive Scale)**.** *For any RNN satisfying the hypotheses of Theorem 4.1, $\sigma > 0$.*

Hence, no choice of recurrent weights W_h—orthogonal, unitary, sparse, or regularized—can prevent diffusive dispersion in saturating RNNs. Setting $\lambda_1 = 0$ converts exponential decay or explosion into symmetric dispersion.

Corollary 4.3 (Diffusive Degradation and Usable Horizon)**.** *Under the hypotheses of Theorem 4.1, for any non-zero $v \in \mathbb{R}^{d_h}$. Then:*

(i) ***Equiprobable Failure:*** *If $\lambda_1 = 0$, then for any polynomial $p(m)$ with > 0 leading coefficient,*

$$\lim_{m\to\infty} \mathbb{P}\left(\frac{\|J_{t\to t+m}v\|}{\|v\|} > p(m)\right) = \tfrac{1}{2} \quad \textit{and} \quad \lim_{m\to\infty} \mathbb{P}\left(\frac{\|J_{t\to t+m}v\|}{\|v\|} < \frac{1}{p(m)}\right) = \tfrac{1}{2}. \tag{10}$$

(ii) ***Usable Horizon:*** *For dynamic range tolerance $\kappa > 1$ and confidence level $1-\alpha$, let $z_{1-\alpha/2} := \Phi^{-1}(1-\alpha/2)$ denote the standard normal quantile. The usable horizon $m^*(\kappa)$—the maximum horizon such that $\mathbb{P}(R_m \in [\kappa^{-1}, \kappa]) \geq 1-\alpha$ to leading order—is the unique positive root of*

$$|\lambda_1|\, m \;+\; z_{1-\alpha/2}\, \sigma \sqrt{m} \;=\; \log \kappa, \tag{11}$$

which reduces to $m^* = \big(\log \kappa / (z_{1-\alpha/2}\sigma)\big)^2$ when $\lambda_1 = 0$.

Equation 11 reveals that gradient transport is jointly constrained by (λ_1, σ). Near criticality ($|\lambda_1| \approx 0$), diffusion σ is the primary bottleneck (Figure 1a). Out of criticality, even if σ is small, $|\lambda_1| \gg 0$ will cause a sharp decline in m^*, as demonstrated by orthogonal recurrences in Section 5.

5 Experiments

The experiments test three predictions made in Sections 3–4: that the limit results (cf. Theorems 3.1 and 4.1) hold, within and beyond the architecture they are proved for; that at matched drift, the diffusion σ still limits training, so it needs to be handled; and that gradient propagation is governed by the usable horizon m^* relative to the target task horizon H_{task}. Vanilla RNNs are the simplest instance of the framework, and the tasks are chosen to isolate gradient transport. Throughout our experiments, $\kappa = 10$ and $\alpha = 0.05$, so $z_{1-\alpha/2} = 1.96$.

5.1 Empirical Validation of the Limit Theorems

Central limit theorem. The spread of $\log \|J_{t\to t+m} v\|$ grows as $\sigma\sqrt{m}$, as predicted by Theorem 4.1 (Figure 2; slope $\approx \frac{1}{2}$), both at $d_u = 1$ and at $d_u = d_h = 128$, with i.i.d inputs, where every hypothesis holds. The prediction involves no fitted parameter, since σ is estimated on separate trajectories. Trained networks show the same agreement (observed-to-predicted ratio 0.99). Correlated sequential MNIST (psMNIST; **?**) inputs steepen the slope to 0.66, but the growth remains diffusive. The invariance of λ_1 asserted by Theorem 3.1 is verified in Appendix E.1.

Extension to other recurrent architectures. The pair (λ_1, σ) is determined by the specific stationary recurrence. We measure it on seven untrained cells calibrated to a common drift, $\lambda_1 \approx -0.01$ (Figure 3a, Appendix E.3). The cells whose recurrent Jacobian depends on the input (vanilla, orthogonal, GRU, LSTM, antisymmetric, selective State Space Models (SSMs)) all have $\sigma > 0$ and a spread growing as $\sqrt{m}$ (slopes 0.46–0.50). At equal drift, σ spans two orders of magnitude, from 0.0011 (selective SSM) to 0.17 (vanilla), and the usable horizon m^* ranges from 33 to 227 steps under a common drift cap $\log\kappa/|\lambda_1| \approx 230$. Even the orthogonal cell, with bounded one-step Jacobians, has $\sigma = 0.16$. The time-invariant SSM, whose Jacobian is constant, has $\sigma = 0$. The diffusion is therefore determined by how the input enters the recurrent path, not by the drift.

5.2 Variance-Aware Regularization at Matched Drift

We add σ^2 to the flossing penalty,

$$\mathcal{R}(\vartheta) \;=\; \alpha \sum_{i \leq k} \lambda_i(\vartheta)^2 \;+\; \beta\, \sigma(\vartheta)^2, \tag{12}$$

estimating σ^2 by batch means (Flegal & Jones, 2010) over the per-step log-stretches $\log R_t[1,1]$ that the differentiable QR recursion already produces (Engelken, 2023; Appendix D.8). The estimator is strongly consistent under the geometric ergodicity of Corollary D.9, and adds negligible computation: $+0.5$ to 3.4% wall-clock over flossing, against 41–103% for the QR recursion itself. Unless stated otherwise we train vanilla RNNs ($d_h = 128$, $\tanh$, Adam, clipping at 1) with $k = 5$, $\alpha = 1$ and $\beta = 30$, and re-measure (λ_1, σ) after training; m^* uses the seed-mean values (Appendix F).

psMNIST. Table 1 compares the variance penalty with flossing over a range of strengths α. On top of flossing, $\beta\sigma^2$ reduces σ threefold ($0.116 \to 0.038$), nearly quadruples m^* ($49 \to 189$), and raises accuracy from 0.834 to 0.860, a gain comparable to that of flossing over the unregularized network. Because the penalty also halves $|\lambda_1|$, the flossing rows serve as a matched-drift control. At $\alpha = 30$,

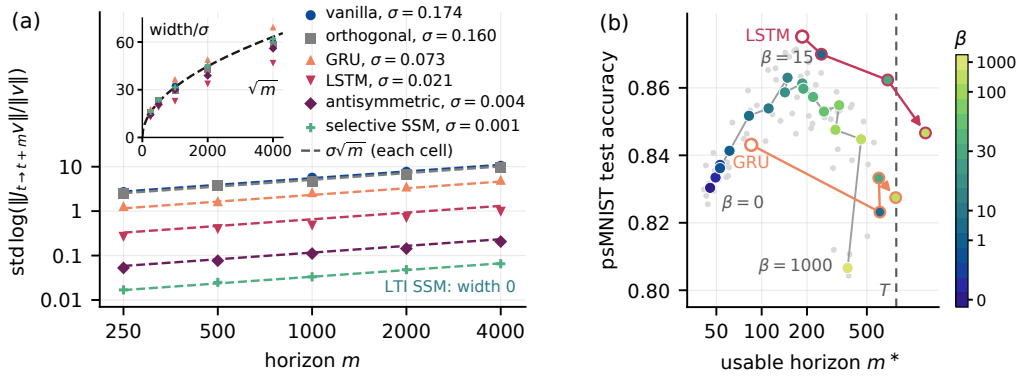

Figure 3: **Diffusion across cells and the effect of the variance penalty.** *(a)* Six untrained cells calibrated to $\lambda_1 \approx -0.01$. Markers: spread of $\log \|J_{t\to t+m}v\|$ over 512 trajectories; dashed: $\sigma\sqrt{m}$, with σ (legend) measured in a separate run. Inset: spread divided by σ. The spread grows as $\sqrt{m}$ in every cell (slopes 0.46–0.50). The time-invariant SSM has $\sigma = 0$ and zero spread. *(b)* psMNIST test accuracy against m^* for the vanilla RNN as β increases. Accuracy peaks at $m^* \approx 150$–190 ($\beta = 15$–30), about a fifth of T (dashed), and drops to 0.807 at $\beta = 1000$. LSTM and GRU: no regularizer (open), then flossing $+\beta\sigma^2$ at $\beta = 3, 30, 300$.

Table 1: **Variance penalty and flossing on psMNIST (vanilla RNN).** Test accuracy (mean ± s.e.m.; 5 seeds, 3 for $\alpha \geq 10$) and (λ_1, σ) measured after training. At matched drift (flossing at $\alpha = 30$ against flossing $+\beta\sigma^2$), flossing retains twice the diffusion and is 3.4 points less accurate.

	accuracy	λ_1	σ	m^*
orthogonal	0.794 ± 0.004	-0.056	0.066	29
unregularized	0.811 ± 0.003	-0.024	0.146	31
flossing, $\alpha = 100$	0.585 ± 0.128	$\mathbf{-0.0052}$	0.090	101
flossing, $\alpha = 1$	0.834 ± 0.002	-0.0149	0.116	49
flossing, $\alpha = 10$	0.826 ± 0.008	-0.0097	0.093	75
flossing, $\alpha = 30$	0.826 ± 0.009	-0.0071	0.078	105
flossing $+\beta\sigma^2$	$\mathbf{0.860 \pm 0.005}$	-0.0067	**0.038**	**189**

flossing reaches the same drift as the variance-aware network ($\lambda_1 = -0.0071$ against -0.0067) but retains twice the diffusion ($\sigma = 0.078$) and is 3.4 points less accurate (Welch $t = 3.4$). Flossing alone does not reduce σ below about 0.08, and training degrades for $\alpha \gtrsim 100$. The orthogonal baseline illustrates the opposite failure: low diffusion, strong contraction, and $m^* = 29$.

Adding problem. On the adding task (Hochreiter & Schmidhuber, 1997), flossing alone already centers the drift: at $T \in \{200, 300\}$ the flossed networks accumulate at most 0.5 nats of drift over the sequence, so the remaining limitation is diffusion. Adding $\beta\sigma^2$ lowers the MSE in 4–6 of 8 paired seeds at every T. At $T = 300$, $\beta = 3$ halves the MSE ($0.036 \to 0.017$) and solves $8/8$ seeds against $5/8$ (paired $t = -2.1$). At $T = 400$ the larger weight $\beta = 30$ is needed ($0.068 \to 0.029$, $6/8$ seeds solved against $3/8$, paired $t = -2.0$); at $T = 200$ the gain is small (Table 2).

5.3 Usable Horizon and Task Horizon

Optimal intermediate penalty. By Proposition 4.2, σ cannot be eliminated, only lowered. Sweeping β from 0 to 1000 reveals an intermediate sweet spot rather than monotonic improvement as $\beta \to \infty$ (Figure 3b). Accuracy peaks at $m^* \approx 150$–190 ($\beta = 15$–30). At $\beta = 1000$, it falls below the unregularized network, and m^* decreases to 372 because the drift contracts faster than σ decreases. The optimum lies well short of the sequence length $T = 784$: psMNIST does not require gradient transport over the full horizon, and suppressing variance beyond what the task demands unnecessarily constrains network expressivity. Gated cells show the same effect. A plain LSTM already operates near the vanilla optimum ($m^* = 186$, accuracy 0.875); the penalty raises its m^* to 250–1200 and lowers its accuracy to 0.847. Similarly, $\beta = 300$ prevents learning on the adding problem at $T \geq 300$, and the best β grows with the horizon, from 3 at $T = 300$ to 30 at $T = 400$.

Table 2: **Effect of the variance penalty across task horizons.** Adding: geometric-mean test MSE over 8 seeds (seeds solved, MSE < 0.05). nsCIFAR of length T (32 image rows followed by $T - 32$ noise steps): test accuracy of a vanilla RNN, 3 seeds. The last two columns add $\beta\sigma^2$ to flossing; †chance level (0.10).

task	T	flossing	$\beta = 3$	$\beta = 30$
adding	200	0.014 (7/8)	0.011 (7/8)	0.015 (8/8)
	300	0.036 (5/8)	**0.017 (8/8)**	0.028 (6/8)
	400	0.068 (3/8)	0.052 (4/8)	**0.029 (6/8)**
nsCIFAR	132	0.240	0.233	0.250
	332	$0.114^\dagger$	0.139	**0.170**
	1000	$0.115^\dagger$	$0.111^\dagger$	$0.094^\dagger$

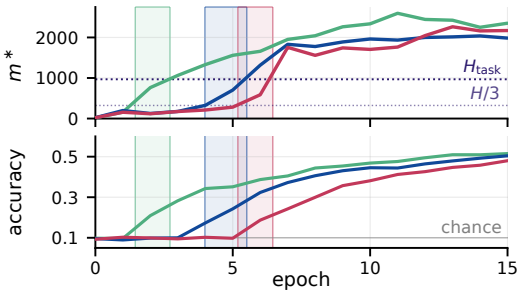

Figure 4: **Accuracy leaves chance as m^* approaches H_{task}.** nsCIFAR, flossed GRU, three seeds (colors). Shaded: epochs over which m^* rises from $H_{\text{task}}/3$ to H_{task} (top); acc. leaves chance within them (bottom).

A necessary condition on the usable horizon. Corollary 4.3 suggests a condition consistent with all of the above. Over the horizon H_{task} required by a task, both the drift $|\lambda_1|H$ and the diffusion $z\sigma\sqrt{H}$ must remain within $\log \kappa$; equivalently, $m^* \geq H_{\text{task}}$. We treat this as a graded necessary condition, not as a threshold. Noise-padded CIFAR-10 (nsCIFAR: 32 rows of pixels followed by Gaussian noise) fixes H_{task} by construction, and the padding length varies it at a fixed cell and dataset (Table 2). Three regimes appear: (i) with 100 noise steps, flossing alone suffices; (ii) with 300, the flossed vanilla RNN is limited by diffusion (7.1 nats), and $\beta\sigma^2$ restores learning ($0.114 \to 0.170$, $t = 4.2$, all seeds improve); the penalty also reduces the drift here ($7.3 \to 1.5$ nats), which the matched-drift control of Table 1 accounts for; and (iii) with the standard 968 steps, the vanilla RNN is limited by drift and no setting learns. The adding problem exhibits the same pattern.

Other recurrent cells. On standard nsCIFAR, the same condition holds across seven cells and 75 runs (Appendix F). Every run that learns keeps the drift near zero and the diffusion below 0.75 nats over 968 steps. Flossing raises the GRU from chance to 0.563 by centering its drift; $\beta\sigma^2$ adds nothing further, since its m^* already exceeds the task horizon ninefold. Beyond the condition, accuracy is determined by the cell (GRU 0.56, chrono-initialized LSTM 0.52, antisymmetric 0.37, selective SSM 0.36). Measured after the first epoch, m^*/H_{task} separates the 42 runs that learn from the 33 that remain at chance-level accuracy (AUC 0.996; Figure 4). Since these failures are drift-limited, λ_1 alone performs comparably (0.98).

6 Discussion

Implications for other architectures. Although Theorem 4.1 is stated for vanilla RNNs, the $\sqrt{m}$ growth of the fluctuations is observed in every cell of Section 5.1 whose recurrent Jacobian depends on the input, including the gated, orthogonal, antisymmetric and selective ones, with σ estimated separately and no fitted parameter (Figure 3a). The pair (λ_1, σ), and hence m^*, is therefore a valid metric of gradient propagation for these models as well, and we expect the limit theorems to extend to them. The two quantities are affected by different design choices. Pointwise constraints such as orthogonality bound each one-step stretch but not its fluctuation (Appendix D.6), whereas σ is determined by how the input enters the recurrent path. In selective SSMs, the diffusion is small, but since the gains satisfy $a_t \in (0, 1]$, $\lambda_1 = 0$ is only attained when $a_t \equiv 1$ on the slowest coordinate. That coordinate then evolves as a time-invariant integrator, with $\sigma = 0$ as for the LTI model (Appendix E.3).

Limitations and open problems. Theorem 4.1 requires $d_u \geq d_h$ and i.i.d. inputs. Figure 2 confirms the prediction in that regime; the trained networks, with $d_u = 1$ and correlated inputs, are covered only empirically. The tolerance κ in m^* is not calibrated. A vanilla RNN solves the adding problem at $T = 100$ with $m^* \approx 14$–30, and the antisymmetric cell learns nsCIFAR with $m^* \approx 630$–$750 < 968$, which may be due to adaptive step sizes and clipping tolerating attenuations of 10^2–10^3 (App. F). Finally, both penalties differentiate through $\vartheta \mapsto (\lambda_1, \sigma)$, which we have shown to be well defined but not differentiable; Appendix E.5 conjectures local Hölder continuity of λ_1 through a numerical test.

REPRODUCIBILITY STATEMENT

All proofs, including required assumptions, are presented in Appendices C and D. Remarks C.4 and D.2 discuss the weight hypotheses and what they exclude. Architectures, hyperparameters, estimator settings, task definitions, seeds and compute are in Appendix F. Every number in Section 5 is read from `runs/*/results.csv` by the same plotting modules that draw the figures, with per-figure sidecar JSONs, so the prose and the panels cannot drift apart. Code, run records and the scripts that produce every figure and table are available at `https://anonymous.4open.science/r/Drift-is-not-enough-BD67/README.md`.

THE USE OF LARGE LANGUAGE MODELS

A large language model was used as a writing and editing aid — condensing prose, checking internal consistency between the text, the tables and the run records, and drafting figure captions — as a coding assistant for experiments, plotting and diagnostic scripts, for help in literature search and for aid in developing and writing theorems, proofs, experimental designs and claims. All LLM-assisted drafts were rewritten and edited by the authors, and every paragraph, proof and experiment result reported was verified by the authors.

REFERENCES

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1120–1128, 2016.

Ludwig Arnold. *Random Dynamical Systems*. Springer Monographs in Mathematics. Springer, Berlin, 1998.

Benny Avelin and Anders Karlsson. Deep limits and a cut-off phenomenon for neural networks. *Journal of Machine Learning Research*, 23(191):1–29, 2022.

Nils Bertschinger and Thomas Natschläger. Real-time computation at the edge of chaos in recurrent neural networks. *Neural Computation*, 16(7):1413–1436, 2004.

Vivek S Borkar. Exploding and vanishing gradients in deep neural networks: the effect of residual connections. *arXiv preprint arXiv:2606.17013*, 2026.

Philippe Bougerol and Jean Lacroix. *Products of Random Matrices with Applications to Schrödinger Operators*, volume 8 of *Progress in Probability and Statistics*. Birkhäuser, Boston, MA, 1985.

Bo Chang, Minmin Chen, Eldad Haber, and Ed H. Chi. AntisymmetricRNN: A dynamical system view on recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Minmin Chen, Jeffrey Pennington, and Samuel S. Schoenholz. Dynamical isometry and a mean field theory of RNNs: Gating enables signal propagation in recurrent neural networks. In *International Conference on Machine Learning (ICML)*, 2018.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

Pedro Duarte and Silvius Klein. *Lyapunov Exponents of Linear Cocycles: Continuity via Large Deviations*, volume 3 of *Atlantis Studies in Dynamical Systems*. Atlantis Press, Paris, 2016.

Rainer Engelken. Gradient flossing: Improving gradient descent through dynamic control of jacobians. *Advances in Neural Information Processing Systems*, 36:10412–10439, 2023.

Rainer Engelken, Fred Wolf, and L. F. Abbott. Lyapunov spectra of chaotic recurrent neural networks. *Physical Review Research*, 5(4):043044, 2023.

N. Benjamin Erichson, Omri Azencot, Alejandro Queiruga, Liam Hodgkinson, and Michael W. Mahoney. Lipschitz recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2021.

James M. Flegal and Galin L. Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2):1034–1070, 2010.

Harry Furstenberg and Harry Kesten. Products of random matrices. *The Annals of Mathematical Statistics*, 31(2):457–469, 1960.

Lyudmila Grigoryeva and Juan-Pablo Ortega. Echo state networks are universal. *Neural Networks*, 108:495–508, 2018.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Ankit Gupta, Karan Goel, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems*, 2022.

Martin Hairer. Ergodic properties of Markov processes, 2018. URL `https://www.hairer.org/notes/Markov.pdf`. Lecture notes.

Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020.

Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled Cayley transform. In *International Conference on Machine Learning (ICML)*, pp. 1969–1978, 2018.

Mikael Henaff, Arthur Szlam, and Yann LeCun. Recurrent orthogonal networks and long-memory tasks. In *International Conference on Machine Learning (ICML)*, pp. 2034–2042, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Li Jing, Yichen Shen, Tena Dubček, John Peurifoy, Scott Skirlo, Yann LeCun, Max Tegmark, and Marin Soljačić. Tunable efficient unitary neural networks (EUNN) and their application to RNNs. In *International Conference on Machine Learning (ICML)*, pp. 1733–1741, 2017.

Rudolf Emil Kalman. Mathematical description of linear dynamical systems. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1(2):152–192, 1963.

Giancarlo Kerg, Kyle Goyette, Maximilian Puelma Touzel, Gauthier Gidel, Eugene Vorontsov, Yoshua Bengio, and Guillaume Lajoie. Non-normal recurrent neural network (nnRNN): Learning long time dependencies while improving expressivity with transient dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Émile Le Page. Théorèmes limites pour les produits de matrices aléatoires. In *Probability Measures on Groups (Oberwolfach, 1981)*, volume 928 of *Lecture Notes in Mathematics*, pp. 258–303. Springer, Berlin, 1982.

Robert Legenstein and Wolfgang Maass. Edge of chaos and prediction of computational performance for neural circuit models. *Neural Networks*, 20(3):323–334, 2007.

G. Manjunath and Herbert Jaeger. Echo state property linked to an input: Exploring a fundamental characteristic of recurrent neural networks. *Neural Computation*, 25(3):671–696, 2013.

Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition, 2009.

Jonas M. Mikhaeil, Zahra Monfared, and Daniel Durstewitz. On the difficulty of learning chaotic dynamics with RNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

A. Emin Orhan and Xaq Pitkow. Improved memory in recurrent neural networks with sequential non-normal dynamics. In *International Conference on Learning Representations (ICLR)*, 2020.

Juan-Pablo Ortega and Florian Rossmannek. Stochastic dynamics learning with state-space systems, 2025. To appear in Mathematical Models and Methods in Applied Sciences.

Valery I. Oseledets. A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems. *Transactions of the Moscow Mathematical Society*, 19:197–231, 1968.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: Theory and practice. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

David Ruelle. Analyticity properties of the characteristic exponents of random matrix products. *Advances in Mathematics*, 32(1):68–80, 1979.

T. Konstantin Rusch and Siddhartha Mishra. Coupled oscillatory recurrent neural network (coRNN): An accurate and (gradient) stable architecture for learning long time dependencies. In *International Conference on Learning Representations (ICLR)*, 2021.

T. Konstantin Rusch, Siddhartha Mishra, N. Benjamin Erichson, and Michael W. Mahoney. Long expressive memory for sequence modeling. In *International Conference on Learning Representations*, 2022.

Jannis Schuecker, Sven Goedeke, and Moritz Helias. Optimal sequence memory in driven random networks. *Physical Review X*, 8(4):041029, 2018.

L. Storm, H. Linander, J. Bec, K. Gustavsson, and B. Mehlig. Finite-time Lyapunov exponents of deep neural networks. *Physical Review Letters*, 132(5):057301, 2024.

Corentin Tallec and Yann Ollivier. Can recurrent neural networks warp time? In *International Conference on Learning Representations*, 2018.

Adrian Trapletti, Friedrich Leisch, and Kurt Hornik. Stationary and integrated autoregressive neural network processes. *Neural Computation*, 12(10):2427–2450, 2000.

Marcelo Viana. *Lectures on Lyapunov Exponents*, volume 145 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2014.

Marcelo Viana and Krerley Oliveira. *Foundations of Ergodic Theory*, volume 151 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2016.

Ryan Vogt, Maximilian Puelma Touzel, Eli Shlizerman, and Guillaume Lajoie. On lyapunov exponents for rnns: Understanding information propagation using dynamical systems tools. *Frontiers in Applied Mathematics and Statistics*, 8:818799, 2022.

Ryan Vogt, Yang Zheng, and Eli Shlizerman. Lyapunov-guided representation of recurrent neural network performance. *Neural Computing and Applications*, 2024. doi: 10.1007/s00521-024-09824-6.

Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, and Chris Pal. On orthogonality and learning recurrent networks with long term dependencies. In *International Conference on Machine Learning (ICML)*, pp. 3570–3578, 2017.

Gilles Wainrib and Mathieu N. Galtier. A local echo state property through the largest Lyapunov exponent. *Neural Networks*, 76:39–45, 2016.

A Extended related work, and terminology

Ergodicity of input-driven recurrent models. The reservoir computing community has studied the input-driven case for two decades under the heading of the echo state property. Manjunath & Jaeger (2013) formulate the property relative to a given input; Wainrib & Galtier (2016) give a local version through the largest Lyapunov exponent; Grigoryeva & Ortega (2018) and, most relevantly, Ortega & Rossmannek (2025) develop fading memory and stochastic echo states as attractor dynamics on the space of probability measures. Our invariant measure π is a stochastic echo state in their sense. Independently, the nonlinear time-series literature established geometric ergodicity for autoregressive neural network processes using precisely the Meyn–Tweedie apparatus collected in Appendix B.3 (Trapletti et al., 2000); our Proposition C.2 is the compact-state-space specialization, where Doeblin's condition replaces a drift argument, and we claim it as a short self-contained proof for the case at hand rather than as a new contribution. The contribution is Theorem 3.1, which combines it with the multiplicative ergodic theorem to obtain the *whole spectrum of the gradient cocycle* as a deterministic function of the weights — a statement about $\partial h_{t+m}/\partial h_t$, not about the state process, and one that the works above do not make.

Products of random matrices in deep networks. Hanin & Nica (2020) and Pennington et al. (2017); Chen et al. (2018) study the singular spectrum of long Jacobian products by random-matrix and mean-field methods. *Dynamical isometry*, which asks that the whole singular spectrum concentrate near one, is strictly stronger than $\lambda_1 = 0$ and than the pointwise constraints of Section 1.1, and gives the third point of the ordering in Appendix D.6. Storm et al. (2024) study finite-time Lyapunov exponents of deep networks.

Terminology. The object we study appears under several names. The physics and reservoir-computing communities call the exponents of an input-driven system *conditional Lyapunov exponents*; Vogt et al. (2022) call them *stochastic Lyapunov exponents*; in the language of Arnold (1998) the pair (driving chain, Jacobian) is a *random dynamical system* and the exponents are those of its linearized cocycle. We use "Lyapunov spectrum of the gradient dynamics" throughout and record the synonyms once so that the result is findable from any of these directions. We reserve *gradient flossing* for the specific method of Engelken (2023) and *Lyapunov regularization* for the family.

B Background

Throughout, P is a transition kernel on a measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and P^m its m-step iterate.

B.1 Ergodic theory and products of random matrices

Definition B.1 (Measure-preserving system)**.** Let $(\Omega, \mathcal{B}, \mu)$ be a probability space and $\theta : \Omega \to \Omega$ measurable. The map θ is *measure preserving* if $\mu(\theta^{-1}B) = \mu(B)$ for all $B \in \mathcal{B}$. A set $B \in \mathcal{B}$ is *invariant* if $\mu(\theta^{-1}B \triangle B) = 0$, and $(\Omega, \mathcal{B}, \mu, \theta)$ is *ergodic* if every invariant set has $\mu(B) \in \{0, 1\}$.

Proposition B.2 (Viana & Oliveira, 2016, Ch. 4)**.** *A measure-preserving system $(\Omega, \mathcal{B}, \mu, \theta)$ is ergodic if and only if every θ-invariant measurable function $f : \Omega \to [-\infty, \infty]$ is constant μ-almost everywhere.*

For reference, the two statements Section 2 describes in prose are the following.

Definition B.3 (Linear cocycle)**.** Let θ preserve a probability measure μ on $(\Omega, \mathcal{B})$ and let $\mathbf{A} : \Omega \to \mathbb{R}^{d\times d}$ be measurable. The *linear cocycle generated by* $\mathbf{A}$ *over* θ is $\mathcal{A}_m(\omega) := \mathbf{A}(\theta^{m-1}\omega) \cdots \mathbf{A}(\omega)$, $m \geq 1$, which obeys $\mathcal{A}_{m+n}(\omega) = \mathcal{A}_n(\theta^m\omega)\, \mathcal{A}_m(\omega)$.

Theorem B.4 (Multiplicative ergodic theorem; Furstenberg & Kesten, 1960; Oseledets, 1968; Viana, 2014)**.** *Let θ preserve μ and let $\log^+ \|\mathbf{A}\| \in L^1(\mu)$. Then there exist measurable, θ-invariant functions $\lambda_1 \geq \cdots \geq \lambda_d : \Omega \to [-\infty, \infty)$ such that, for μ-a.e. ω,*

$$\lim_{m\to\infty} \tfrac{1}{m} \log \sigma_i\big(\mathcal{A}_m(\omega)\big) \;=\; \lambda_i(\omega), \qquad i = 1, \ldots, d. \tag{13}$$

Corollary B.5. *Suppose in addition that $(\Omega, \mathcal{B}, \mu, \theta)$ is ergodic. Then each λ_i is μ-a.e. equal to a constant, and moreover $\lambda_1 = \inf_{m\geq 1} \frac{1}{m} \mathbb{E}_\mu \log \|\mathcal{A}_m\|$.*

Theorem B.4 is the conjunction of two classical results. Furstenberg–Kesten (Furstenberg & Kesten, 1960) states that if $\log^+ \|\mathbf{A}\| \in L^1(\mu)$ then $\lambda_1(\omega) = \lim_m \frac{1}{m} \log \|\mathcal{A}_m(\omega)\|$ exists in $[-\infty, \infty)$ for μ-a.e. ω and is θ-invariant; it follows from Kingman's subadditive ergodic theorem applied to $\omega \mapsto \log \|\mathcal{A}_m(\omega)\|$, and subadditivity is also what yields the variational expression in Corollary B.5. The extension from the top exponent to the full list of singular value growth rates equation 13 is the one-sided Oseledets theorem (Oseledets, 1968); see Viana (2014, Ch. 3–4) for both, in the form used here.

Neither result assumes ergodicity, and neither concludes that the λ_i are constants: what they give is θ-invariance, i.e. measurability with respect to the invariant σ-algebra $\mathcal{I} = \{B \in \mathcal{B} : \theta^{-1}B = B \text{ up to } \mu\text{-null sets}\}$. In general λ_i is therefore a random variable, constant on each ergodic component of μ but free to differ between components; the decomposition of μ into ergodic components makes this precise (Viana & Oliveira, 2016, Ch. 5).

Proof of Corollary B.5. By Theorem B.4 each λ_i is θ-invariant, so Proposition B.2 makes it μ-a.e. equal to a constant. For the variational expression, Kingman's subadditive ergodic theorem applied to $f_m(\omega) = \log \|\mathcal{A}_m(\omega)\|$, which satisfies $f_{m+n} \le f_n \circ \theta^m + f_m$ and $f_1^+ \in L^1(\mu)$, gives $\mathbb{E}_\mu \lambda_1 = \inf_{m \ge 1} \frac{1}{m} \mathbb{E}_\mu f_m$; since λ_1 is a.e. constant, $\lambda_1 = \mathbb{E}_\mu \lambda_1$. □

B.2 Markov chains: Doeblin's condition and limit theorems

Theorem B.6 (Doeblin; Hairer, 2018, Thm. 4.29, Meyn & Tweedie, 2009, Thm. 16.2.4)**.** *Suppose there are $\alpha > 0$ and a probability measure η on $\mathcal{X}$ with $P(x, \cdot) \ge \alpha\, \eta(\cdot)$ for every $x \in \mathcal{X}$. Then P admits a unique invariant probability measure π, and $\sup_{B \in \mathcal{B}(\mathcal{X})} |P^m(x, B) - \pi(B)| \le (1 - \alpha)^m$ for all $x \in \mathcal{X}$, $m \ge 1$.*

Theorem B.7 (Hairer, 2018, Thm. 5.7)**.** *The set of invariant probability measures of P is convex, and its extreme points are exactly those π for which the stationary path measure $\mathbb{P}_\pi$ on $\mathcal{X}^{\mathbb{N}}$ is ergodic for the forward shift. In particular, if P has exactly one invariant probability measure π, then π is extreme and $\mathbb{P}_\pi$ is shift-ergodic.*

Section 4 needs the k-step version of Theorem B.6 and the limit theorems it unlocks. We collect them here in the form used; all are from Meyn & Tweedie (2009, Ch. 16–17). Recall that $\mathcal{X}$ is *ν_k-small* if $P^k(x, \cdot) \ge \nu_k(\cdot)$ for every $x \in \mathcal{X}$, for a non-trivial measure ν_k.

Theorem B.8 (k-step Doeblin; Meyn & Tweedie, 2009, Thm. 16.0.2 and eq. (16.11))**.** *Suppose there are $k \ge 1$, $\delta \in (0, 1]$ and a probability measure ϱ on $\mathcal{X}$ with*

$$P^k(x, \cdot) \;\ge\; \delta\, \varrho(\cdot) \qquad \text{for every } x \in \mathcal{X}, \tag{14}$$

i.e. $\mathcal{X}$ is itself small. Then P is uniformly ergodic*: it is ψ-irreducible, aperiodic and positive Harris recurrent, it has a unique invariant probability measure π, and $\sup_{x \in \mathcal{X}} \|P^m(x, \cdot) - \pi\|_{\mathrm{TV}} \le 2(1 - \delta)^{\lfloor m/k \rfloor}$ for every $m \ge 1$.*

For $V : \mathcal{X} \to [1, \infty)$ put $\|\mu\|_V := \sup_{|g| \le V} |\mu(g)|$. The chain is *$V$-uniformly ergodic* if $\sup_x \|P^n(x, \cdot) - \pi\|_V / V(x) \to 0$ as $n \to \infty$; uniform ergodicity is the case $V \equiv 1$.

Theorem B.9 (Meyn & Tweedie, 2009, Thm. 16.0.1)**.** *Let Φ be ψ-irreducible and aperiodic and let $V \ge 1$. Then Φ is V-uniformly ergodic if and only if the drift condition $PV \le (1 - \beta)V + b\,\mathbf{1}_C$ holds for some $\beta \in (0, 1)$, $b < \infty$, some petite set C, and some V_0 equivalent to V.*

Corollary B.10 (Bounded drift)**.** *Suppose $\mathcal{X}$ is small in the sense of equation 14 and $V \ge 1$ satisfies $C_0 := \sup_x PV(x) < \infty$. Then Φ is V-uniformly ergodic and $\pi(V) \le C_0$.*

Proof. By Theorem B.8 the chain is ψ-irreducible and aperiodic, and $\mathcal{X}$ is small, hence petite, hence so is every subset of it. Put $\beta = \frac{1}{2}$, $C := \{V \le 2C_0\}$ and $b := C_0$. If $V(x) > 2C_0$ then $PV(x) \le C_0 < \frac{1}{2}V(x) = (1 - \beta)V(x)$; if $x \in C$ then $PV(x) \le C_0 = b \le (1 - \beta)V(x) + b$. Theorem B.9 applies with $V_0 = V$. Finally $\pi(V) = \pi(PV) \le C_0$ by invariance and Tonelli. □

Theorem B.11 (Limit theorems for additive functionals; Meyn & Tweedie, 2009, Thm. 17.0.1)**.** *Let Φ be positive Harris with invariant probability π, let $g : \mathcal{X} \to \mathbb{R}$ be measurable, write $\bar{g} := g - \pi(g)$ and $S_m(g) := \sum_{n=1}^m g(\Phi_n)$.*

(i) (LLN) *If* $\pi(|g|) < \infty$ *then* $\frac{1}{m} S_m(g) \to \pi(g)$ *almost surely, from every initial condition.*

(ii) (Asymptotic variance) *If* Φ *is* V*-uniformly ergodic and* $g^2 \le V$*, then*

$$\gamma_g^2 \;:=\; \mathbb{E}_\pi\big[\bar{g}(\Phi_0)^2\big] \;+\; 2\sum_{n\ge 1} \mathbb{E}_\pi\big[\bar{g}(\Phi_0)\,\bar{g}(\Phi_n)\big] \tag{15}$$

is well defined, finite and non-negative, and ***coincides with the asymptotic variance***

$$\lim_m \tfrac{1}{m}\mathbb{E}_\pi[S_m(\bar{g})^2].$$

(iii) Under the hypotheses of (ii), if $\gamma_g^2 = 0$ *then* $m^{-1/2} S_m(\bar{g}) \to 0$ *a.s.*

(iv) (CLT and LIL) *Under the hypotheses of (ii), if* $\gamma_g^2 > 0$ *then, from every initial condition,* $m^{-1/2} S_m(\bar{g}) \xrightarrow{d} \mathcal{N}(0, \gamma_g^2)$ *and, almost surely,*

$$\limsup_{m\to\infty} \frac{S_m(\bar{g})}{\sqrt{2\gamma_g^2\, m \log\log m}} = 1, \qquad \liminf_{m\to\infty} \frac{S_m(\bar{g})}{\sqrt{2\gamma_g^2\, m \log\log m}} = -1.$$

Theorem B.12 (Poisson equation and martingale decomposition; Meyn & Tweedie, 2009, Thm. 17.4.3, Thm. 17.4.4)**.** *Let* Φ *be positive Harris and let* $\hat{g}$ *solve Poisson's equation* $\hat{g} - P\hat{g} = g - \pi(g)$ *with* $\pi(|\hat{g}|) < \infty$*. Then, under* $\mathbb{P}_\pi$*,*

$$S_m(\bar{g}) \;=\; M_m \;+\; P\hat{g}(\Phi_0) - P\hat{g}(\Phi_m), \qquad M_m := \sum_{n=1}^{m} \big[\hat{g}(\Phi_n) - P\hat{g}(\Phi_{n-1})\big], \tag{16}$$

and M_m *is a martingale for the natural filtration. If in addition* $\pi(\hat{g}^2) < \infty$ *then the constant of Theorem B.11(ii) is* $\gamma_g^2 = \pi\big(\hat{g}^2 - (P\hat{g})^2\big) = \mathbb{E}_\pi\big[(\hat{g}(\Phi_1) - P\hat{g}(\Phi_0))^2\big]$.

Two remarks on how these are used. First, Theorem B.11 concerns $g(\Phi_n)$, a function of the state alone, whereas the quantity of interest in Section 4 is the one-step log-stretching, which depends on the state *and* on the fresh input. The standard remedy, used in Appendix D, is to enlarge the state so that the summand becomes a function of the enlarged state. Second, g there is allowed to be unbounded, at the price of exhibiting a V with $g^2 \le V$; Corollary B.10 is what makes this cheap in our setting, where the enlarged chain forgets its state in one step.

B.3 Markov chains: the Meyn–Tweedie apparatus

The results below are what Proposition C.2 runs on, via Assumption 3 of Section 2, proved in Appendix C.4. They are also the route to a recurrence on an unbounded state space, which would in addition need the drift theory of Theorem B.17.

Definition B.13 (Nonlinear state space model)**.** $\{X_k\}$ is an $\mathrm{NSS}(F)$ model if $X_k = F(X_{k-1}, W_k)$ for a smooth $F : \mathcal{X} \times \mathcal{O}_w \to \mathcal{X}$ and i.i.d. disturbances $\{W_k\}$ with law Γ admitting a lower semicontinuous density γ_w; the *control set* is the open set $\mathcal{O}_w := \{w : \gamma_w(w) > 0\}$. The associated deterministic *control model* $\mathrm{CM}(F)$ is *forward accessible* if for every x_0 the attainable set $A_+(x_0) := \bigcup_{k\ge 0}\{F_k(x_0, w_1, \ldots, w_k) : w_i \in \mathcal{O}_w\}$ has non-empty interior.

Proposition B.14 (Meyn & Tweedie, 2009, Prop. 7.1.5)**.** *If an* $\mathrm{NSS}(F)$ *model has a lower semicontinuous disturbance density and its control model* $\mathrm{CM}(F)$ *is forward accessible, then* $\mathcal{X}$ *is a union of open small sets and the chain is a* T*-chain.*

Proposition B.15 (Meyn & Tweedie, 2009, Prop. 6.2.1)**.** *A* T*-chain possessing a reachable point* x^* *is* ψ*-irreducible with* $\psi = T(x^*, \cdot)$.

Theorem B.16 (Meyn & Tweedie, 2009, Thm. 6.2.5, Thm. 5.5.7)**.** *For a* ψ*-irreducible* T*-chain every compact set is petite; if the chain is moreover aperiodic, every petite set is small.*

Theorem B.17 (Foster–Lyapunov drift; Meyn & Tweedie, 2009, Thm. 1.3.1, Thm. 11.3.4)**.** *Suppose there exist a small set* C*, a constant* $b < \infty$ *and a non-negative* V*, finite everywhere, with* $\int P(x, \mathrm{d}y) V(y) \le V(x) - 1 + b\,\mathbf{1}_C(x)$ *for all* x*. Then the chain admits a unique invariant probability measure.*

C Deferred proofs for Section 3

This appendix proves Theorem 3.1. We first outline the argument (Appendix C.1), then state the two propositions that Section 3 describes in prose (Appendix C.2), and finally give the proofs.

C.1 Proof roadmap for Section 3

The result we want to use is the Multiplicative Ergodic Theorem (MET) in its ergodic form (Theorem B.4 and Corollary B.5, (Meyn & Tweedie, 2009)): for a linear cocycle over an ergodic measure-preserving system with $\log^+ \|\mathbf{A}\|$ integrable, the Lyapunov exponents are almost surely constant. Working backwards from this statement, the proof has to supply three ingredients, shown in Figure 5.

1. A cocycle whose products are the gradient Jacobians (Proposition C.1). The one-step Jacobian $J_t = D_t W_h$ depends on the previous state h_{t-1} *and* on the fresh input u_t, so it is not a function of the hidden-state chain alone. Pairing the two into $z_t := (h_{t-1}, u_t)$ gives a Markov chain on $\mathcal{Z}$ with $J_t = A(z_t)$, and the long-term Jacobian $J_{t\to t+m}$ is the cocycle $\mathcal{A}_m$ evaluated along the shifted path $\theta^t\omega$. Since $0 < \phi' \le 1$, we have $\|A\| \le \|W_h\|$, so the integrability hypothesis holds automatically.

2. An ergodic base (Proposition C.2). By Theorem B.7, the stationary path measure is ergodic as soon as the chain $\{z_t\}$ has a unique invariant law. The input coordinate of z_{t+1} is drawn afresh and independently at every step, so this reduces to uniqueness for the hidden-state chain, and on the compact cube $\mathcal{X}$ uniqueness follows from an ℓ-step Doeblin minorisation (Theorem B.8). When $d_u < d_h$ the one-step kernel has no density, so the minorisation cannot be read off a transition density. We obtain it from the Meyn–Tweedie apparatus of Appendix B.3 instead, with each clause of Assumption 3 doing one job (bottom row of Figure 5): trajectory controllability (a) makes the chain a T-chain; the reachable equilibrium (b) makes it ψ-irreducible and aperiodic; and compactness of $\mathcal{X}$ then makes the whole state space small.

3. Transfer to a fixed initial state and time origin (Lemma C.6). The Multiplicative Ergodic Theorem holds for μ-almost every path, that is, for the chain started from its stationary law, whereas a network is run from a fixed h_0. Convergence of $P^m(x_0, \cdot)$ to π in total variation, supplied by Proposition C.2(ii), transfers every shift-invariant event of full μ-measure to the chain started at x_0. The event on which the exponents converge is shift-invariant because every one-step Jacobian is invertible, and the same invariance gives the statement for all time origins t at once.

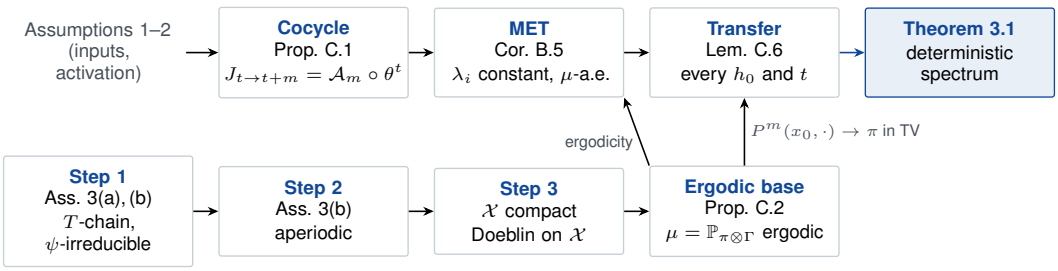

Figure 5: **Structure of the proof of Theorem 3.1.** *Top row:* the gradient product is a linear cocycle over the input-driven Markov chain (Proposition C.1); over an ergodic base the Multiplicative Ergodic Theorem (MET) makes its exponents constant for μ-almost every path; Lemma C.6 extends this to every initial state and every time origin. *Bottom row:* the ergodic base (Proposition C.2) is obtained from the Meyn–Tweedie apparatus, each step using one clause of Assumption 3 or the compactness of $\mathcal{X}$.

Only the *existence* of the minorisation equation 20 is used. Its constants come out of the Meyn–Tweedie theory and are not explicit, and no rate of convergence enters the proof (Remark C.5). The minorisation concerns the hidden state only; Appendix D needs a minorisation for the pair (state, transported direction), which does not follow from Proposition C.2 and is constructed separately in Appendix D.4.

C.2 THE COCYCLE FORMULATION AND THE ERGODIC BASE

Proposition C.1 (Cocycle formulation)**.** *Let Assumptions 1–2 hold, and set*

$$\mathcal{Z} := \mathcal{X}\times\mathbb{R}^{d_u}, \qquad z_t := (h_{t-1},\, u_t) \in \mathcal{Z}, \qquad A(h,u) := \operatorname{diag}\big(\phi'(W_h h+W_u u+b_h)\big)\, W_h. \tag{17}$$

Then $\{z_t\}_{t\ge 1}$ is a Markov chain on $\mathcal{Z}$ with kernel $Q((h,u),\cdot) = \delta_{\phi(W_h h+W_u u+b_h)}\otimes\Gamma$. Let $\Omega := \mathcal{Z}^{\mathbb{N}}$ carry the product σ-algebra, let θ be the forward shift, let $\mathbf{A}(\omega) := A(\omega_1)$, and let μ be any shift-invariant law of $\{z_t\}$ on Ω. Then $\mathcal{A}_m$ is a linear cocycle over $(\Omega,\mathcal{B},\mu,\theta)$ with $\|\mathbf{A}\| \le \|W_h\|$ pointwise, and

$$\mathcal{A}_m(\theta^t\omega) \;=\; J_{t\to t+m} \qquad \textit{for all } t\ge 0,\ m\ge 1. \tag{18}$$

Because the kernel Q draws the input coordinate afresh at every step, producing a shift-invariant ergodic μ reduces to a uniqueness statement for the hidden-state chain

$$h_t \;=\; F(h_{t-1},u_t), \qquad F(h,u) := \phi(W_h h + W_u u + b_h), \qquad h_t \in \mathcal{X}, \tag{19}$$

whose transition kernel we denote by P.

Proposition C.2 (Ergodicity of the driving process)**.** *Let Assumptions 1, 2 and 3 hold. Then:*

(i) there are $n_0 \ge 1$, $\delta \in (0,1]$ and a probability measure ϱ on $\mathcal{X}$ with

$$P^{n_0}(x,B) \;\ge\; \delta\,\varrho(B) \qquad \textit{for all } x\in\mathcal{X},\ B\in\mathcal{B}(\mathcal{X}); \tag{20}$$

(ii) consequently, the chain equation 19 has a unique invariant probability measure π, and for every $x\in\mathcal{X}$ and $m\ge 1$, $\sup_B |P^m(x,B)-\pi(B)| \le (1-\delta)^{\lfloor m/n_0\rfloor}$;

(iii) the chain $\{z_t\}$ has the unique invariant law $\pi\otimes\Gamma$, so the stationary path measure $\mu := \mathbb{P}_{\pi\otimes\Gamma}$ on Ω is θ-invariant and ergodic.

C.3 PROOF OF PROPOSITION C.1

Proof. Markov property. With F as in equation 19, $z_{t+1} = (h_t, u_{t+1}) = \big(F(z_t),\, u_{t+1}\big)$. The first component is a deterministic function of z_t, and the second is independent of $z_1,\dots,z_t$ with law Γ; this is the stated kernel. Comparing the definition of A in equation 17 with $J_t = \operatorname{diag}(\phi'(a_t))W_h$ gives $A(z_t) = A(h_{t-1},u_t) = J_t$.

Cocycle identity. By construction $\mathbf{A}(\theta^{m-1}\omega) = A(\omega_m)$, so along a trajectory $\omega = (z_1,z_2,\dots)$ we have $\mathbf{A}(\theta^{m-1}\omega) = A(z_m) = J_m$ and hence, by equation 4,

$$\mathcal{A}_m(\omega) \;=\; J_m J_{m-1}\cdots J_1 \;=\; J_{0\to m}.$$

Applying this to the shifted trajectory $\theta^t\omega = (z_{t+1},z_{t+2},\dots)$ gives $\mathcal{A}_m(\theta^t\omega) = J_{t+m}\cdots J_{t+1} = J_{t\to t+m}$, which is equation 18. Shift-invariance of μ makes θ measure preserving, so $\mathcal{A}_m$ is a linear cocycle in the sense of Definition B.3.

Uniform bound. Since $0 < \phi' \le 1$, the diagonal factor is a contraction and $\|A(h,u)\| \le \|\operatorname{diag}(\phi'(a))\|\,\|W_h\| \le \|W_h\|$. Hence $\log^+\|\mathbf{A}\| \le \log^+\|W_h\|$, a finite constant, so $\log^+\|\mathbf{A}\| \in L^\infty(\mu)\subset L^1(\mu)$ for every probability measure μ on Ω. □

C.4 PROOF OF PROPOSITION C.2

When $d_u < d_h$, the one-step kernel $P(x,\cdot)$ is carried by the image of the map $u\mapsto F(x,u)$, which has dimension at most d_u, so $P(x,\cdot)$ has no density and a minorisation cannot be read off a lower bound on one. We use the Meyn–Tweedie apparatus of Appendix B.3 instead. It yields minorisations on compact sets, and compactness of $\mathcal{X}$ then makes the minorisation uniform over the whole state space, as equation 20 requires.

The Meyn–Tweedie framework for nonlinear state space models is stated on open state spaces, so we regard equation 19 as a chain on $\mathbb{R}^{d_h}$ rather than on $\mathcal{X}$. By Assumption 3, F is C^∞ on $\mathbb{R}^{d_h}\times\mathbb{R}^{d_u}$, and it maps into $(-1,1)^{d_h}\subset\mathcal{X}$, so the chain enters $\mathcal{X}$ after one step and never leaves it. With the input law Γ of Assumption 1, whose density is strictly positive and lower semicontinuous, this is an $\mathrm{NSS}(F)$ model in the sense of Definition B.13 with control set $\mathcal{O}_w = \mathbb{R}^{d_u}$.

Lemma C.3 (Forward accessibility)**.** *Under Assumption 3,* $\mathrm{CM}(F)$ *is forward accessible.*

Proof. Write $\Psi_x^{(\ell)}(u_1, \dots, u_\ell) := h_\ell$ for the ℓ-step map started at $h_0 = x$. Differentiating equation 3 along the recursion gives $\partial h_\ell / \partial u_\ell = D_\ell W_u$ and, for $j < \ell$,

$$\frac{\partial h_\ell}{\partial u_j} \;=\; D_\ell W_h \, D_{\ell-1} W_h \cdots D_{j+1} W_h \; D_j W_u, \qquad D_i = \mathrm{diag}(\phi'(a_i)),$$

so that, ordering the blocks $j = \ell, \ell-1, \dots, 1$,

$$D\Psi_x^{(\ell)}(u) \;=\; D_\ell \big[\; W_u \;\; W_h D_{\ell-1} W_u \;\; \cdots \;\; W_h D_{\ell-1} W_h \cdots D_1 W_u \;\big]. \tag{21}$$

Each D_i is a positive diagonal matrix because $\phi' > 0$, so D_ℓ is invertible and the rank of equation 21 equals the rank in equation 6. Assumption 3(a) therefore states that, for every $x \in \mathcal{X}$, the map $\Psi_x^{(\ell)}$ is a submersion at some control. By the submersion theorem it is open near that control, so the attainable set $A_+(x)$ contains a non-empty open set. For $x \in \mathbb{R}^{d_h} \setminus \mathcal{X}$, one step lands in some $x' \in (-1, 1)^{d_h}$, and $A_+(x) \supseteq A_+(x')$ has non-empty interior by the previous sentence. □

Remark C.4 (On Assumption 3). The bracket in equation 6 has ℓd_u columns, so clause (a) is available for every $d_u \geq 1$ once $\ell \geq \lceil d_h / d_u \rceil$; at $\ell = 1$ it reads $\mathrm{rank}\, W_u = d_h$, the full-rank condition of Assumption 4. Clause (a) does not follow from the classical Kalman condition $\mathrm{rank}[\, W_u \;\; W_h W_u \;\; \cdots \;\; W_h^{d_h - 1} W_u \,] = d_h$, which is the same expression evaluated at the unattainable configuration $D_i = I$: the D_i are the saturation profile that the control itself realizes. For each fixed x, however, the set of controls at which the rank in equation 6 equals d_h is open, and for real-analytic ϕ such as $\tanh$ its complement is either every control or a Lebesgue-null set, since the sum of squared maximal minors is real-analytic in the control. Clause (b) is an irreducibility condition: it excludes state spaces that split into non-communicating regions, as in Trapletti et al. (2000). Equilibria themselves always exist, by Brouwer's theorem applied to $x \mapsto \phi(W_h x + W_u u^\star + b_h)$ on $\mathcal{X}$.

Proof of Proposition C.2. Step 1: a ψ-irreducible T-chain. By Lemma C.3 and Proposition B.14, $\{h_t\}$ is a T-chain and $\mathbb{R}^{d_h}$ is a union of open small sets. We next show that $h^\star$ is a *reachable* point, that is, $\sum_{n \geq 1} P^n(x, O) > 0$ for every open $O \ni h^\star$ and every state x (Meyn & Tweedie, 2009, §6.1.2). Fix such O and x. One step carries x into $\mathcal{X}$, and by Assumption 3(b) some control $(u_1, \dots, u_n)$ then steers the state into O. The n-step map is continuous in the controls, so every control in an open neighbourhood of $(u_1, \dots, u_n)$ also steers the state into O, and this neighbourhood has positive $\Gamma^{\otimes n}$-measure because $\gamma > 0$. Hence $\sum_n P^n(x, O) > 0$, and Proposition B.15 makes the chain ψ-irreducible. We take ψ maximal.

Step 2: aperiodicity. This step uses the equilibrium of Assumption 3(b). A direct argument is not available: an equilibrium of equation 19 need not be attracting, so a ball around $h^\star$ need not be mapped into itself. Instead we exhibit a small set around $h^\star$ that is small at two consecutive times. Recall that if C is ν-small with $\nu(C) > 0$, the period of the chain is the greatest common divisor of the times n at which C is small with a minorising measure proportional to ν (Meyn & Tweedie, 2009, §5.4.3).

By Step 1, $h^\star$ lies in an open small set O_0: there are $M \geq 1$ and a non-trivial measure ν_M with $P^M(y, \cdot) \geq \nu_M(\cdot)$ for all $y \in O_0$. Fix $\rho > 0$ with $\overline{B}(h^\star, \rho) \subseteq O_0$. The maps $(x, u_1) \mapsto F(x, u_1)$ and $(x, u_1, u_2) \mapsto F\big(F(x, u_1), u_2\big)$ are continuous and take the value $h^\star$ at $(h^\star, u^\star)$ and $(h^\star, u^\star, u^\star)$ respectively, so there are $\rho' \in (0, \rho]$ and $\varepsilon > 0$ such that

$$F(x, u_1) \in B(h^\star, \rho) \quad \text{and} \quad F\big(F(x, u_1), u_2\big) \in B(h^\star, \rho) \tag{22}$$

whenever $\|x - h^\star\| \leq \rho'$ and $\|u_1 - u^\star\|, \|u_2 - u^\star\| \leq \varepsilon$. Put $C := \overline{B}(h^\star, \rho')$ and $\alpha := \Gamma\big(\{u : \|u - u^\star\| \leq \varepsilon\}\big)$, which is strictly positive because γ is. Then $C \subseteq O_0$ is small, being a subset of a small set, and equation 22 gives

$$\inf_{x \in C} P^j\big(x, B(h^\star, \rho)\big) \;\geq\; \alpha^j \qquad \text{for } j = 1, 2. \tag{23}$$

Moreover, C is a neighbourhood of the reachable point $h^\star$, hence accessible from every state, hence $\psi(C) > 0$ (Meyn & Tweedie, 2009, §4.2). By ψ-irreducibility there is $m \geq 0$ with $\nu_M P^m(C) > 0$.

Write $\nu := \nu_M P^m$, so that $\nu(C) > 0$ and $P^{M+m}(y, \cdot) \geq \nu(\cdot)$ for every $y \in O_0$. Combining this with equation 23 and $B(h^\star, \rho) \subseteq O_0$ gives, for $x \in C$ and $j = 1, 2$,

$$P^{M+m+j}(x, \cdot) \;\geq\; \int_{B(h^\star,\rho)} P^j(x, \mathrm{d}y)\, P^{M+m}(y, \cdot) \;\geq\; \alpha^j\, \nu(\cdot).$$

So C is ν-small with $\nu(C) > 0$, and it is small with minorising measures proportional to ν at the two consecutive times $M + m + 1$ and $M + m + 2$. The period therefore divides $\gcd(M + m + 1, M + m + 2) = 1$, and the chain is aperiodic.

Step 3: $\mathcal{X}$ is small. The chain is a ψ-irreducible aperiodic T-chain, so by Theorem B.16 every compact set is petite and every petite set is small. The set $\mathcal{X} = [-1, 1]^{d_h}$ is compact, so there are $n_0 \geq 1$ and a non-trivial measure ν_{n_0} with $P^{n_0}(x, \cdot) \geq \nu_{n_0}(\cdot)$ for every $x \in \mathcal{X}$. Since $P^{n_0}(x, \cdot)$ is carried by $(-1, 1)^{d_h}$, so is ν_{n_0}. Putting $\delta := \nu_{n_0}(\mathcal{X}) \in (0, 1]$ and $\varrho := \nu_{n_0}/\delta$ gives equation 20, which is (i).

Step 4: uniqueness and convergence. Every invariant probability measure of P is carried by $(-1, 1)^{d_h}$, because $\pi = \pi P$ and each $P(x, \cdot)$ is; we may therefore read P as a kernel on $\mathcal{X}$, which the chain enters at time 1 and never leaves. On $\mathcal{X}$, equation 20 is the k-step Doeblin condition equation 14 with $k = n_0$, so Theorem B.8 supplies a unique invariant probability measure π together with $\sup_{x\in\mathcal{X}} \|P^m(x, \cdot) - \pi\|_{\mathrm{TV}} \leq 2(1 - \delta)^{\lfloor m/n_0 \rfloor}$. Since $\|\mu - \nu\|_{\mathrm{TV}} = 2 \sup_B |\mu(B) - \nu(B)|$, this is (ii).

Step 5: the enlarged chain. Let λ be invariant for the kernel Q of Proposition C.1. If $z_t \sim \lambda$, then $z_{t+1} = (F(z_t), u_{t+1})$ has law $\eta \otimes \Gamma$, where η is the law of $F(z_t)$, because u_{t+1} is independent of z_t with law Γ. The law of z_{t+1} is thus a product measure whatever λ is, and invariance forces $\lambda = \eta \otimes \Gamma$. Under this λ the two components are independent with laws η and Γ, so η is the law of $F(h, u)$ with $h \sim \eta$ and $u \sim \Gamma$ independent; that is, $\eta = \eta P$. By (ii), $\eta = \pi$ and $\lambda = \pi \otimes \Gamma$, and the same computation shows that $\pi \otimes \Gamma$ is invariant. Uniqueness makes $\pi \otimes \Gamma$ an extreme point of the set of invariant measures of Q, and Theorem B.7 shows that $\mu = \mathbb{P}_{\pi\otimes\Gamma}$ is θ-invariant and ergodic, which is (iii). □

Remark C.5 (Compactness and the rate). The minorisation equation 20 is uniform over $\mathcal{X}$ only because $\mathcal{X}$ is compact and absorbing. On an unbounded state space, Step 3 would give minorisations on compact sets only, and uniform ergodicity would have to be replaced by V-uniform ergodicity under a Foster–Lyapunov drift condition (Theorems B.17 and B.9). The constants n_0 and δ are not explicit, but no rate is used downstream: Lemma C.6 needs only that $P^m(x_0, \cdot) \to \pi$ in total variation.

C.5 Proof of Theorem 3.1

The Multiplicative Ergodic Theorem holds for μ-almost every path, that is, for the chain started from its stationary law $\pi \otimes \Gamma$, whereas the network is run from a fixed h_0. The following lemma transfers almost-sure statements about shift-invariant events from the stationary chain to every initial state, using only the convergence in total variation of Proposition C.2(ii).

Lemma C.6 (Transfer to arbitrary initialisation)**.** *Let Assumptions 1, 2 and 3 hold, let $\mathbb{P}_{x_0}$ denote the law on Ω of $\{z_t\}_{t\geq 1}$ started from $h_0 = x_0 \in \mathcal{X}$, and let $E \subseteq \Omega$ satisfy $\mu(E) = 1$ and $\theta^{-1}E = E$. Then $\mathbb{P}_{x_0}(E) = 1$ for every $x_0 \in \mathcal{X}$.*

Proof. Write $\varsigma_m := P^m(x_0, \cdot) \otimes \Gamma$ for the law of z_{m+1} under $\mathbb{P}_{x_0}$, and $\varsigma_\infty := \pi \otimes \Gamma$ for its stationary counterpart. Let $\mathbb{P}_z$ denote the law of the chain with $z_1 = z$. By the Markov property,

$$\mathbb{P}_{x_0} \circ \theta^{-m} = \int_{\mathcal{Z}} \mathbb{P}_z(\cdot)\, \varsigma_m(\mathrm{d}z), \qquad \mu = \int_{\mathcal{Z}} \mathbb{P}_z(\cdot)\, \varsigma_\infty(\mathrm{d}z).$$

Since $\mathbb{P}_z(E) \leq 1$, the second identity and $\mu(E) = 1$ force $\varsigma_\infty(G) = 1$ for $G := \{z \in \mathcal{Z} : \mathbb{P}_z(E) = 1\}$. Since $\theta^{-m}E = E$ for every m, the first identity gives

$$\begin{aligned}\mathbb{P}_{x_0}(E) \;&=\; \int_{\mathcal{Z}} \mathbb{P}_z(E)\, \varsigma_m(\mathrm{d}z) \;\geq\; \varsigma_m(G) \;\geq\; \varsigma_\infty(G) - \sup_{B\in\mathcal{B}(\mathcal{Z})} \big|\varsigma_m(B) - \varsigma_\infty(B)\big| \\ &=\; 1 - \sup_{B\in\mathcal{B}(\mathcal{X})} \big|P^m(x_0, B) - \pi(B)\big|,\end{aligned}$$

where the last equality holds because the two measures share the second factor Γ. By Proposition C.2(ii), the right-hand side tends to 1 as $m \to \infty$. □

Proof of Theorem 3.1. By Proposition C.1, $\mathcal{A}_m$ is a linear cocycle over $(\Omega, \mathcal{B}, \mu, \theta)$ with $\log^+ \|\mathbf{A}\|$ bounded, and by Proposition C.2(iii) the base system is ergodic. Theorem B.4 and Corollary B.5 therefore provide constants $\lambda_1 \geq \cdots \geq \lambda_{d_h}$, determined by μ and $\mathbf{A}$, hence by $(W_h, W_u, b_h, \phi, \Gamma)$ alone, such that

$$E := \Big\{\omega \in \Omega : \lim_{m\to\infty} \tfrac{1}{m} \log \sigma_i\big(\mathcal{A}_m(\omega)\big) = \lambda_i, \;\; i = 1, \ldots, d_h\Big\} \tag{24}$$

has $\mu(E) = 1$. The bound $\lambda_1 \leq \log \|W_h\|$ follows from $\|\mathcal{A}_m\| \leq \|W_h\|^m$.

E is shift-invariant. By the cocycle identity, $\mathcal{A}_{m+1}(\omega) = \mathcal{A}_m(\theta\omega)\mathbf{A}(\omega)$. Under Assumptions 2 and 3, the matrix $\mathbf{A}(\omega) = \operatorname{diag}(\phi'(a))W_h$ is invertible for every ω, since $\phi' > 0$ and W_h is nonsingular. Hence, for each i,

$$\sigma_i\big(\mathcal{A}_m(\theta\omega)\big)\, \sigma_{d_h}\big(\mathbf{A}(\omega)\big) \;\leq\; \sigma_i\big(\mathcal{A}_{m+1}(\omega)\big) \;\leq\; \sigma_i\big(\mathcal{A}_m(\theta\omega)\big)\, \sigma_1\big(\mathbf{A}(\omega)\big),$$

with both multiplicative factors in $(0, \infty)$ for each fixed ω. Taking $\frac{1}{m}\log$ and letting $m \to \infty$ shows that the limits in equation 24 exist for ω if and only if they exist for $\theta\omega$, with the same values; that is, $\theta^{-1}E = E$.

Uniformity in ω is not needed here, and it is not available: when Γ has unbounded support, the pre-activation $a = W_h h + W_u u + b_h$ is unbounded and $\phi'(a) \to 0$, so $\sigma_{d_h}(\mathbf{A})$ is not bounded away from zero. The argument uses only that, for each fixed ω, a single factor contributes nothing to the $\frac{1}{m}\log$ limit. Control of this lower tail is what the moment conditions of Assumption 6 provide in Appendix D.

Arbitrary h_0. Lemma C.6 upgrades $\mu(E) = 1$ to $\mathbb{P}_{x_0}(E) = 1$ for every $x_0 \in \mathcal{X}$.

Arbitrary time origin. Fix $t \geq 0$. By equation 18, $J_{t\to t+m} = \mathcal{A}_m(\theta^t\omega)$, so equation 7 for time origin t is the statement $\theta^t\omega \in E$, that is, $\omega \in \theta^{-t}E = E$, which has $\mathbb{P}_{x_0}$-probability one. Since E does not depend on t, a single exceptional null set serves all time origins simultaneously. □

D Deferred material for Section 4

D.1 Proof roadmap

The proofs of Section 4 rest on the limit theorems for additive functionals of Markov chains collected in Theorem B.11 (Meyn & Tweedie, 2009). These theorems require the driving chain to be positive Harris recurrent and V-uniformly ergodic, for some V that dominates the square of the summand. The argument therefore has three steps: we write the log-gradient norm as an additive functional of a suitable chain, reduce both hypotheses to a single minorisation condition, and prove that condition. We outline the steps here and carry them out in the subsections that follow.

1. State Space and Additive Representation (Appendix D.3): The log-scaling $\log(\|J_{0\to m}v_0\|/\|v_0\|)$ cannot be expressed as a function of the hidden state h_t alone, as it couples to the transported direction $[v_t] \in \mathbb{P}^{d_h-1}$ and the incoming input. We therefore adjoin directions to form the *projective chain* $w_n := (h_n, [v_n]) \in \mathcal{W} := \mathcal{X} \times \mathbb{P}^{d_h-1}$ (written $w_n = (h_n, \bar{v}_n)$ when a unit representative $\bar{v}_n = v_n/\|v_n\|$ is convenient), yielding the exact additive representation in equation 30. To place this in standard Markovian form, we define the *enlarged chain* $\Phi_n := (w_{n-1}, a_n) \in \mathcal{Y} := \mathcal{W} \times \mathbb{R}^{d_h}$ (equation 31), so that the one-step log-stretching $f(\Phi_n)$ is an additive functional of the state alone.

2. Positive Harris Recurrence and V-Uniform Ergodicity via Minorization: Theorem B.11 requires the driving chain to be positive Harris recurrent and V-uniformly ergodic with $f^2 \leq V$. By Theorem B.8, a k-step Doeblin minorization on the entire state space ($P^k(x, \cdot) \geq \delta\varrho(\cdot)$) simultaneously delivers ψ-irreducibility, aperiodicity, and positive Harris recurrence. Once the space is small, Corollary B.10 guarantees that a one-step bounded drift condition ($\sup_{\mathcal{Y}} PV < \infty$) yields V-uniform ergodicity for the drift function $V(w, a) := 1 + 2C_1^2 + 2L(a)^2 \geq f(w, a)^2$ (Lemma D.5). Both properties therefore hinge entirely on constructing a uniform minorization condition.

3. Constructing the Minorization via Saturation (Appendix D.4): Unlike the hidden state chain, directions cannot be mixed by noise alone without exploiting the activation geometry. Under Assumption 5, saturating activations permit directional control: driving all pre-activations except one to large values concentrates $D(a)W_h$ near rank one, collapsing all incoming directions onto a coordinate axis $[e_i]$ (*collapse*, Lemma D.6). Full-rank input noise then spreads this ray into a fixed open ball $B_i \subset \mathbb{P}^{d_h-1}$ with strictly positive density (*spread*, Lemma D.7). As schematized in Figure 6, alternating collapse and spread funnels all directions into a common coordinate ray $[e_1]$ in 3 steps, and generates a joint product density over directions and pre-activations in steps 4–5 (Proposition D.8). This minorization transfers directly to the enlarged chain $\mathcal{Y}$ in 6 steps (Corollary D.9).

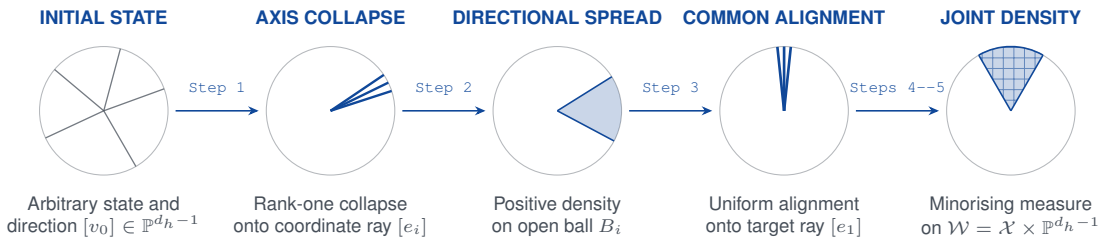

Figure 6: **Five-step projective minorisation (Proposition D.8).** The construction drives arbitrary directions into a common minorising measure on $\mathcal{W} = \mathcal{X} \times \mathbb{P}^{d_h-1}$. Steps 1–3 bring an arbitrary initial direction $[v_0]$ close to the coordinate ray $[e_1]$ by alternating saturation-induced contraction (collapse, Lemma D.6) with noise-driven dispersion (spread, Lemma D.7). Steps 4–5 generate a product density $\mathrm{vol}|_{B_1} \otimes \mathrm{Leb}|_C$ over the direction and the next pre-activation, whose image under the one-step update is the minorising measure of the uniform Doeblin condition on $\mathcal{W}$.

4. Consequences of the minorisation. With the hypotheses of Theorem B.11 in place, the rest of the argument has three parts.

- **Asymptotic Drift and Vector CLT (Appendix D.5):** Proposition D.10 identifies the stationary drift $\pi(f)$ with the deterministic top Lyapunov exponent λ_1. Lemma D.12 constructs a *bounded* solution ψ to Poisson's equation via geometric mixing, producing a square-integrable martingale difference decomposition. Proposition D.14 then invokes Theorem B.11 to establish the Central Limit Theorem and Law of the Iterated Logarithm along any fixed non-zero vector v, which Theorem 4.1 generalizes across all initial states and time origins.
- **Strict Non-Degeneracy of Diffusion (Appendix D.6):** Proposition 4.2 proves that the asymptotic variance satisfies $\sigma > 0$ strictly for any choice of recurrent weights W_h. If $\sigma = 0$ were to hold, the log-stretching would be cohomologous to a constant with the bounded corrector ψ; rescaling pre-activations under saturating non-linearities shifts the stretching by $\log \varepsilon$ while preserving the projective state, directly contradicting the boundedness of ψ.
- **Diffusive Degradation and Usable Horizon (Appendix D.7):** Corollary 4.3 derives the practical consequences: driving $\lambda_1 \to 0$ renders gradient explosion and vanishing equiprobable with superpolynomial dispersion, and the usable horizon $m^*(\kappa)$ is governed jointly by the drift-diffusion relation $|\lambda_1| m + z_{1-\alpha/2}\sigma\sqrt{m} = \log \kappa$.

D.2 Further Assumptions

Assumption 5 (Saturation profile)**.** The activation derivative ϕ' is even and continuously differentiable, with $\phi'(0) = 1$, $\phi'' < 0$ on $(0, \infty)$, and $\lim_{a\to\infty} \phi'(a) = 0$. Consequently ϕ' maps $[0, \infty)$ decreasingly and bijectively onto $(0, 1]$, and restricts to a C^1 diffeomorphism of $(0, \infty)$ onto $(0, 1)$. We write $(\phi')^{-1} : (0, 1] \to [0, \infty)$ for the inverse of $\phi'|_{[0,\infty)}$.

Assumption 6 (Input and saturation moments)**.** There exists $\epsilon > 0$ such that $\mathbb{E}\|u_1\|^{2+\epsilon} < \infty$ and

$$\sup_{h\in\mathcal{X}} \mathbb{E}\big|\log \phi'\big((W_h h + W_u u_1 + b_h)_i\big)\big|^{2+\epsilon} \;<\; \infty, \qquad i = 1, \ldots, d_h. \tag{25}$$

Assumption 7 (Dense recurrent weights)**.** Every entry of the recurrent weight matrix is non-zero: $(W_h)_{ij} \neq 0$ for all $i, j \in \{1, \ldots, d_h\}$.

Section 4 replaces the ℓ-step controllability of Assumption 3 by the full-rank input condition of Assumption 4, which gives the pre-activation a density on all of $\mathbb{R}^{d_h}$ (Lemma D.4); Remark D.2

explains why the proofs below use it. Since Assumption 3 also asks for $\phi \in C^\infty$, which is not assumed here, we first show that Theorem 3.1 remains available.

Lemma D.1 (Well-posedness under full-rank input control)**.** *Under Assumptions 1, 2 and 4, the conclusions of Proposition C.2, with $n_0 = 1$, and of Theorem 3.1 hold.*

Proof. Let C be the closed unit ball of $\mathbb{R}^{d_h}$, and let q and $q_C > 0$ be as in Lemma D.4 below, which uses only Assumptions 1 and 4. For the kernel P of the hidden-state chain equation 19, every $x \in \mathcal{X}$ and every Borel $B \subseteq \mathcal{X}$,

$$P(x, B) \;=\; \int_{\mathbb{R}^{d_h}} q(x, a)\, \mathbf{1}_B\big(\phi(a)\big)\, \mathrm{d}a \;\geq\; q_C\, \mathrm{Leb}\big(\{a \in C : \phi(a) \in B\}\big),$$

which is equation 20 with $n_0 = 1$, $\delta := q_C\, \mathrm{Leb}(C)$ and $\varrho := \phi_*(\mathrm{Leb}|_C)/\mathrm{Leb}(C)$. The proof of Proposition C.2 derives (ii) and (iii) from (i) alone (Step 4 and part (iii)), and the proof of Theorem 3.1, including Lemma C.6, uses Assumption 3 only through Proposition C.2 and the nonsingularity of W_h, which Assumption 4 also provides. □

Remark D.2 (On the requirement $d_u \geq d_h$). Assumption 4 is a hypothesis of our proof technique, not a condition we expect the conclusions to need. It enters only through Lemma D.4, which gives the pre-activation a density on all of $\mathbb{R}^{d_h}$; both the one-step collapse of Lemma D.6 and the dilation argument for $\sigma > 0$ in Appendix D.6 rely on it. When $d_u < d_h$, the pre-activation stays within distance $\|W_h\|\sqrt{d_h} + \|b_h\|$ of the d_u-dimensional subspace $\mathrm{range}\, W_u$, so these single-step configurations are no longer available. Contraction can instead accumulate over several steps, as in Assumption 3. We therefore expect Theorem 4.1 and Proposition 4.2 to extend to this regime; a proof would require a multi-step minorization of the projective chain, which we do not pursue here. The numerical evidence is consistent with this. With $d_u = 1$ and i.i.d. inputs, the empirical width grows as $m^{0.52}$ and matches the independently measured $\sigma\sqrt{m}$ (Figure 2b, Appendix E.2), and every trained network in Table 1, all with $d_u = 1$, has $\sigma > 0$.

Verification for standard architectures. Assumptions 5–7 hold generically:

- **Activation:** For $\phi(x) = \tanh(x)$, $\phi'(x) = \mathrm{sech}^2(x)$ satisfies $\phi'(0) = 1$, $\phi''(x) = -2\tanh(x)\,\mathrm{sech}^2(x) < 0$ for $x > 0$, and $\phi'(x) \to 0$ as $x \to \infty$, satisfying Assumption 5.
- **Moments:** As $|x| \to \infty$, $\log \mathrm{sech}^2(x) = -2|x| + 2\log 2 + o(1)$, so $|\log \phi'(x)| \leq C_\phi(1 + |x|)$ for a universal constant $C_\phi < \infty$. Since $\mathcal{X} = [-1, 1]^{d_h}$ is compact, $|(W_h h + W_u u_1 + b_h)_i| \leq \|W_h\|_\infty + \|b_h\|_\infty + \|W_u\|_\infty \|u_1\|$. Thus, condition equation 25 reduces directly to $\mathbb{E}\|u_1\|^{2+\epsilon} < \infty$, holding for any sub-Gaussian, Gaussian, or finite-moment input distribution.
- **Connectivity:** Dense recurrent weights satisfy Assumption 7 Lebesgue-almost everywhere on $\mathbb{R}^{d_h \times d_h}$.

Technical roles in the proof pipeline.

- **Contraction via saturation (Assumption 5):** Driving $|a_j| \to \infty$ for $j \neq i$ while maintaining $|a_i| \leq 1$ concentrates $D(a)W_h$ arbitrarily close to rank one with range $\mathbb{R}e_i$, collapsing all incoming directions onto coordinate ray $[e_i]$ (Lemma D.6). Additionally, the diffeomorphism property provides the coordinate scaling map S_ε that establishes $\sigma > 0$ (Proposition 4.2).
- **Uniform integrability and drift (Assumption 6):** While $f(w, a) = \log \|D(a)W_h \bar{v}\|$ is bounded above by $\log \|W_h\|$, it is unbounded below under saturation. Assumption 6 guarantees that the Lyapunov drift function $V(w, a) = 1 + 2C_1^2 + 2L(a)^2 \geq f(w, a)^2$ satisfies $\sup_{\mathcal{Y}} PV < \infty$, unlocking V-uniform ergodicity and square-integrability of the Poisson solution (Lemma D.5, Corollary D.9).
- **Regularity of the projective chart (Assumption 7):** Ensuring $(W_h)_{ij} \neq 0$ guarantees that the normalized column vectors $\xi^i := W_h e_i / \|W_h e_i\|$ avoid all projective coordinate hyperplanes ($\xi^i_j \neq 0$ for all j). This ensures the directional pushforward under $D(a)W_h$ lands in an open affine coordinate chart with non-vanishing Jacobian determinant, generating the uniform directional spread density (Lemma D.7).

Notation and structural constants. Throughout the rest of this appendix, Assumptions 1, 2, 4, and 5–7 are in force, and we use the following notation.

- $\mathbb{P}^{d_h-1}$ is the real projective space of $\mathbb{R}^{d_h}$, with the angular metric $\mathrm{d}([u],[v]) := \arccos \frac{|\langle u,v\rangle|}{\|u\|\|v\|}$ and the normalised Riemannian volume vol.
- $D(a) := \mathrm{diag}(\phi'(a))$ is the activation Jacobian at the pre-activation $a \in \mathbb{R}^{d_h}$.
- The structural constants of the recurrent matrix are

$$c_0 \;:=\; \frac{\sigma_{\min}(W_h)}{\sqrt{d_h}} \;>\; 0, \qquad w_{\min} \;:=\; \min_{i,j} |(W_h)_{ij}| \;>\; 0, \tag{26}$$

  where $c_0 > 0$ because W_h is nonsingular (Assumption 4) and $w_{\min} > 0$ by Assumption 7.

Since $\|x\|_\infty \geq \|x\|_2/\sqrt{d_h}$ for every $x \in \mathbb{R}^{d_h}$, every unit vector $\bar{v} \in \mathbb{S}^{d_h-1}$ satisfies

$$\max_{i=1,\dots,d_h} \big|(W_h\bar{v})_i\big| \;\geq\; \frac{1}{\sqrt{d_h}}\|W_h\bar{v}\|_2 \;\geq\; \frac{\sigma_{\min}(W_h)}{\sqrt{d_h}} \;=\; c_0. \tag{27}$$

Thus at least one coordinate of $W_h\bar{v}$ is bounded away from zero, uniformly over directions. This is the coordinate along which the first collapse in the proof of Proposition D.8 is performed.

D.3 The projective chain and the enlarged chain

To describe gradient transport along individual sample paths, we track the hidden state and the transported direction jointly. Fix a non-zero vector $v_0 \in \mathbb{R}^{d_h}$ and define the transported vectors $v_n := J_n v_{n-1} = D(a_n)W_h v_{n-1}$ for $n \geq 1$. Because W_h is nonsingular (Assumption 4) and $\phi' > 0$ (Assumption 2), each one-step Jacobian $D(a_n)W_h$ is invertible, so $v_n \neq 0$ for all n.

Let $[v_n] := \mathbb{R}v_n \in \mathbb{P}^{d_h-1}$ be the direction of v_n. The joint process

$$w_n \;:=\; \big(h_n, [v_n]\big) \in \mathcal{W} \;:=\; \mathcal{X} \times \mathbb{P}^{d_h-1}, \qquad n \geq 0, \tag{28}$$

is a time-homogeneous Markov chain on $\mathcal{W}$, driven by the pre-activations $a_n = W_h h_{n-1} + W_u u_n + b_h$; we denote its transition kernel by $\bar{Q}$. We write a state of $\mathcal{W}$ as $w = (h, \bar{v})$, where $\bar{v} \in \mathbb{S}^{d_h-1}$ is a unit representative of the direction; every expression below that involves $\bar{v}$ is invariant under $\bar{v} \mapsto -\bar{v}$. The chain evolves as $w_n = \Theta(w_{n-1}, a_n)$ under the deterministic update map $\Theta : \mathcal{W} \times \mathbb{R}^{d_h} \to \mathcal{W}$,

$$\Theta\big((h,[v]),a\big) \;:=\; \Big(\phi(a),\, \big[D(a)W_h\bar{v}\big]\Big). \tag{29}$$

Note that Θ does not depend on its hidden-state argument: given the pre-activation, the next state does not involve h_{n-1}. Where convenient we therefore regard Θ as a map $\mathbb{P}^{d_h-1} \times \mathbb{R}^{d_h} \to \mathcal{W}$.

Telescoping the one-step norms $\|v_n\|/\|v_{n-1}\|$ gives the additive representation

$$\log \frac{\|J_{0\to m}v_0\|}{\|v_0\|} \;=\; \sum_{n=1}^{m} f(w_{n-1}, a_n), \qquad f\big((h,[v]),a\big) \;:=\; \log\big\|D(a)W_h\bar{v}\big\|, \tag{30}$$

where f is a well-defined Borel function on $\mathcal{W} \times \mathbb{R}^{d_h}$ because $\|D(a)W_h(-\bar{v})\| = \|D(a)W_h\bar{v}\|$.

Each summand $f(w_{n-1}, a_n)$ depends on both the current state w_{n-1} and the incoming pre-activation a_n. We therefore define the enlarged Markov chain

$$\Phi_n \;:=\; (w_{n-1}, a_n) \in \mathcal{Y} \;:=\; \mathcal{W} \times \mathbb{R}^{d_h}, \qquad n \geq 1, \tag{31}$$

with natural filtration $\mathcal{F}_n := \sigma(w_0, a_1, \dots, a_n)$ for $n \geq 1$ and $\mathcal{F}_0 := \sigma(w_0)$. The chain $\{\Phi_n\}_{n\geq 1}$ is time-homogeneous, with transition kernel P: given $\Phi_n = (w_{n-1}, a_n)$, the next state $w_n = \Theta(w_{n-1}, a_n)$ is determined, and the next pre-activation $a_{n+1} = W_h\phi(a_n) + W_u u_{n+1} + b_h$ depends only on Φ_n and on the input $u_{n+1} \sim \Gamma$, which is independent of $\mathcal{F}_n$. The log-gradient norm ratio is therefore exactly an additive functional of the enlarged chain,

$$\log \frac{\|J_{0\to m}v_0\|}{\|v_0\|} \;=\; \sum_{n=1}^{m} f(\Phi_n) \;=:\; S_m(f). \tag{32}$$

Writing $f_n := f(\Phi_n)$ for the n-th summand, the constant of Theorem 4.1 is the *asymptotic* variance of the per-step log-stretching rather than its plain variance:

$$\sigma^2 \;=\; \mathrm{Var}\,(f_1) \;+\; 2\sum_{n\ge 1}\mathrm{Cov}\,(f_1, f_{1+n}), \qquad f_n \;=\; \log\frac{\|J_{n-1\to n}\,v_{n-1}\|}{\|v_{n-1}\|}, \tag{33}$$

where the moments are taken under the stationary law of the enlarged chain equation 31. The series converges absolutely because that chain is V-uniformly ergodic (Corollary D.9), and Proposition D.14 shows that equation 33 is indeed the variance appearing in equation 9.

The mixing and moment properties of $\{\Phi_n\}$ rest on the regularity of the conditional law of the pre-activation, which full-rank input control provides.

Lemma D.3. *Let Assumptions 1 and 4 hold, and let Γ_u be the law of $W_u u_1$ on $\mathbb{R}^{d_h}$. Then Γ_u has a density g with respect to Lebesgue measure on $\mathbb{R}^{d_h}$ which is strictly positive everywhere and lower semicontinuous.*

Proof. If $d_u = d_h$ then W_u is invertible and $g(v) = |\det W_u|^{-1}\gamma(W_u^{-1}v)$, which inherits positivity and lower semicontinuity from γ. If $d_u > d_h$, let $W_u^\dagger = W_u^\top(W_u W_u^\top)^{-1}$ be the right inverse of W_u and let $N \in \mathbb{R}^{d_u\times(d_u-d_h)}$ have orthonormal columns spanning $\ker W_u$. The map $(v,w)\mapsto W_u^\dagger v + Nw$ is a linear bijection $\mathbb{R}^{d_h}\times\mathbb{R}^{d_u-d_h}\to\mathbb{R}^{d_u}$ with constant Jacobian $c_W := |\det[\,W_u^\dagger \;\; N\,]| > 0$, and $W_u(W_u^\dagger v + Nw) = v$. By the change of variables formula, for every Borel $B\subseteq\mathbb{R}^{d_h}$,

$$\Gamma_u(B) = \int_{\{u:\,W_u u\in B\}}\gamma(u)\,\mathrm{d}u = \int_B \underbrace{c_W\int_{\mathbb{R}^{d_u-d_h}}\gamma\big(W_u^\dagger v + Nw\big)\,\mathrm{d}w}_{=:g(v)}\;\mathrm{d}v.$$

Positivity of g is immediate from $\gamma > 0$. For lower semicontinuity, let $v_n\to v$. Then $\liminf_n \gamma(W_u^\dagger v_n + Nw) \ge \gamma(W_u^\dagger v + Nw)$ for each fixed w, and Fatou's lemma gives $\liminf_n g(v_n)\ge g(v)$. □

Lemma D.4 (Uniform noise density)**.** *The conditional law of a_n given $h_{n-1} = h$ has density $q(h,a) = g(a - W_h h - b_h)$, where g is the density of $W_u u_1$. For every compact $C\subset\mathbb{R}^{d_h}$, $q_C := \inf\{q(h,a) : h\in\mathcal{X},\, a\in C\} > 0$.*

Proof. By Lemma D.3, the law of $W_u u_1$ has a strictly positive, lower semicontinuous density g on $\mathbb{R}^{d_h}$, and q has the stated form because u_n is independent of h_{n-1}. The set $\{a - W_h h - b_h : h\in\mathcal{X},\, a\in C\}$ is the image of a compact set under a continuous map, hence compact, and a strictly positive lower semicontinuous function attains a positive minimum on a compact set. □

Lemma D.5 (Moment bounds and a drift function)**.** *Put $C_1 := |\log\sigma_{\min}(W_h)| + \big|\log\|W_h\|\big|$, $L(a) := \sum_i |\log\phi'(a_i)|$, and*

$$V(w,a) \;:=\; 1 + 2C_1^2 + 2L(a)^2 \;\ge\; 1. \tag{34}$$

Then, for all $w\in\mathcal{W}$ and $a\in\mathbb{R}^{d_h}$,

$$\log\sigma_{\min}(W_h) + \min_i\log\phi'(a_i) \;\le\; f(w,a) \;\le\; \log\|W_h\|, \qquad f(w,a)^2 \le V(w,a).$$

Moreover, $C_0 := \sup_{\mathcal{Y}} PV < \infty$ under Assumption 6, where $PV(w,a) := \mathbb{E}\big[V(\Phi_{n+1}) \mid \Phi_n = (w,a)\big]$ is the one-step expectation under the kernel of the enlarged chain equation 31. In particular, $F(w) := \mathbb{E}[f(w,a_1)\mid w_0 = w]$ is bounded and measurable on $\mathcal{W}$.

Proof. The two-sided bound follows from $\|D(a)W_h\bar v\| \le \max_i\phi'(a_i)\|W_h\| \le \|W_h\|$ and $\|D(a)W_h\bar v\| \ge \min_i\phi'(a_i)\,\sigma_{\min}(W_h)$. Since $\log\phi' \le 0$, we have $|\min_i\log\phi'(a_i)| \le L(a)$, so $|f|\le C_1 + L$ and $f^2\le 2C_1^2 + 2L^2\le V$. For the drift, note that given $\Phi_n = (w,a)$ the next pre-activation has law $q(\phi(a),\cdot)$ with $\phi(a)\in\mathcal{X}$. Using $(\sum_{i=1}^{d_h} x_i)^2 \le d_h\sum_i x_i^2$,

$$\begin{aligned} PV(w,a) \;&=\; 1 + 2C_1^2 + 2\,\mathbb{E}\big[L(a_{n+1})^2 \mid h_n = \phi(a)\big] \\ &\le\; 1 + 2C_1^2 + 2d_h\sum_i \sup_{h\in\mathcal{X}}\mathbb{E}\big|\log\phi'\big((W_h h + W_u u_1 + b_h)_i\big)\big|^2, \end{aligned}$$

which is finite by Assumption 6 and independent of (w,a). Finally, $|F(w)|\le\mathbb{E}[\,|f(w,a_1)|\mid w_0 = w\,]\le C_1 + \mathbb{E}[L(a_1)\mid w_0 = w]$, which is bounded uniformly in w by the same estimate. □

D.4 A MINORISATION CONDITION FOR THE PROJECTIVE CHAIN

This subsection contains the main technical step of the appendix: Proposition C.2 concerns the hidden state and says nothing about directions. The mechanism is that a saturating activation lets the network *steer* directions. Choosing one pre-activation coordinate small and the others large makes $D(a)W_h$ arbitrarily close to a rank-one matrix with range $\mathbb{R}e_i$, which collapses every direction onto $[e_i]$ at once. Alternating collapse and spread drives all directions into a common open set, after which two further steps produce a density jointly in state and direction.

We first fix some notation. For each i, let $\xi^i := W_h e_i/\|W_h e_i\|$, whose coordinates are all non-zero by Assumption 7, and let

$$\mathcal{C}_i := \Big\{[\zeta] \in \mathbb{P}^{d_h-1} :\ \zeta_j \neq 0\ \forall j,\ \ \operatorname{sign}(\zeta_j\zeta_l) = \operatorname{sign}(\xi^i_j\xi^i_l)\ \forall j,l\Big\}$$

be the open sign cell of $[\xi^i]$, a well-defined, non-empty open subset of $\mathbb{P}^{d_h-1}$. Since $\{[\zeta] : (W_h\zeta)_1 = 0\}$ is a projective hyperplane, it cannot contain the open set $\mathcal{C}_i$. We may therefore fix a closed ball $B_i \subset \mathcal{C}_i$ of positive radius that avoids this hyperplane, and, by compactness, a constant $c_1 > 0$ with

$$|(W_h\bar v)_1| \ \geq\ c_1 \qquad \text{for every unit } \bar v \text{ with } [\bar v] \in B_i,\ \text{for every } i. \tag{35}$$

Lemma D.6 (Collapse)**.** *For every $c > 0$ and $\eta \in (0, \pi/8)$ there is $p_A(\eta, c) > 0$ such that for every $h \in \mathcal{X}$, every index i, and every unit vector $\bar v$ with $|(W_h\bar v)_i| \geq c$,*

$$\mathbb{P}\Big(\mathrm{d}\big([D(a_1)W_h\bar v], [e_i]\big) \leq \eta \,\Big|\, h_0 = h\Big)\ \geq\ p_A(\eta, c).$$

Proof. Put $K_i(M) := \{a : |a_i| \leq 1,\ M \leq |a_j| \leq M+1\ (j \neq i)\}$, a compact set whose Lebesgue measure depends on neither i nor M. By Assumption 5, ϕ' is even and strictly decreasing on $[0,\infty)$, so for $a \in K_i(M)$ we have $\phi'(a_i) \geq \phi'(1)$ and $\phi'(a_j) \leq \phi'(M)$ for $j \neq i$. Hence

$$\big|\big(D(a)W_h\bar v\big)_i\big| \geq \phi'(1)\,c, \qquad \big\|\big(D(a)W_h\bar v\big)_{-i}\big\| \leq \sqrt{d_h - 1}\,\phi'(M)\,\|W_h\|,$$

where x_{-i} denotes x with its i-th coordinate removed. Since $\tan \mathrm{d}([x],[e_i]) = \|x_{-i}\|/|x_i|$, this gives $\tan \mathrm{d}([D(a)W_h\bar v],[e_i]) \leq \sqrt{d_h-1}\,\phi'(M)\|W_h\|/(\phi'(1)c)$, which is at most $\tan\eta$ once $M \geq M(\eta, c)$, because $\phi'(M) \to 0$. By Lemma D.4, the event $\{a_1 \in K_i(M)\}$ has conditional probability at least $q_{K_i(M)}\mathrm{Leb}(K_i(M)) > 0$, uniformly in h and $\bar v$. Taking $p_A(\eta, c)$ to be the minimum of this bound over the finitely many indices i makes it independent of i. □

Lemma D.7 (Spread)**.** *For each i there are $\eta_i > 0$ and $\rho_i > 0$ such that for every $h \in \mathcal{X}$ and every unit vector $\bar v$ with $\mathrm{d}([\bar v],[e_i]) \leq \eta_i$, the conditional law of $[D(a_1)W_h\bar v]$ given $h_0 = h$ is bounded below by ρ_i vol on B_i:*

$$\mathbb{P}\big([D(a_1)W_h\bar v] \in A \mid h_0 = h\big)\ \geq\ \rho_i\,\mathrm{vol}(A) \qquad \textit{for every Borel } A \subseteq B_i.$$

Proof. Write $\bar\xi := W_h\bar v/\|W_h\bar v\|$, so that $[D(a)W_h\bar v] = [D(a)\bar\xi]$. The proof has three steps: we place $\bar\xi$ in a compact set of admissible directions, change variables from the pre-activation to a chart of $\mathbb{P}^{d_h-1}$, and bound the resulting density from below uniformly.

Admissible directions. Choose the representative $\bar v$ with $\langle \bar v, e_i\rangle > 0$. Then $\|\bar v - e_i\| \to 0$ as $\mathrm{d}([\bar v],[e_i]) \to 0$, and hence $\bar\xi \to \xi^i$. The coordinates of ξ^i satisfy $|\xi^i_j| \geq w_{\min}/\|W_h e_i\| \geq w_{\min}/\|W_h\|$, so for η_i small enough $\bar\xi$ lies in the compact set

$$\Xi_i := \big\{\xi : \|\xi\| = 1,\ \operatorname{sign}(\xi_j) = \operatorname{sign}(\xi^i_j)\ \forall j,\ |\xi_j| \geq \beta_\xi\big\}, \qquad \beta_\xi := \frac{w_{\min}}{2\|W_h\|} > 0.$$

In particular $[\bar\xi] \in \mathcal{C}_i$. The set Ξ_i depends on neither h nor $\bar v$; all constants below are chosen uniformly over $\xi \in \Xi_i$ and then applied to this particular $\bar\xi$.

Change of variables. Use the affine chart $\chi([\zeta]) := (\zeta_j/\zeta_1)_{j\neq 1}$, a diffeomorphism on $\{\zeta_1 \neq 0\} \supset \mathcal{C}_i$, and put $\bar Y_i := \chi(B_i)$, a compact subset of $\chi(\mathcal{C}_i)$ on which no coordinate vanishes. Restrict the pre-activation a to the open orthant $(0,\infty)^{d_h}$ and change variables to

$$s := \phi'(a_1) \in (0,1), \qquad y_j := \frac{\phi'(a_j)\,\bar\xi_j}{\phi'(a_1)\,\bar\xi_1}\ \ (j \neq 1),$$

so that $\chi([D(a)\bar{\xi}]) = y$. By Assumption 5, ϕ' is a C^1 diffeomorphism of $(0,\infty)$ onto $(0,1)$ with $\phi'' < 0$. The map $T_{\bar{\xi}} : a \mapsto (s, y)$ is therefore injective (s determines a_1, after which y_j determines a_j) and a C^1 diffeomorphism onto its image, with triangular Jacobian determinant $\phi''(a_1)\prod_{j\neq 1}\phi''(a_j)\bar{\xi}_j/(s\bar{\xi}_1) \neq 0$. The same formulas define T_ξ for an arbitrary $\xi \in \Xi_i$.

Uniform lower bound. For $y \in \bar{Y}_i$ and $\xi \in \Xi_i$ set $r_j := y_j\xi_1/\xi_j$. The sign of y_j is that of $\xi_j^i\xi_1^i$, as is the sign of ξ_1/ξ_j, so $r_j > 0$; by compactness of $\bar{Y}_i \times \Xi_i$ there are $0 < r_- \le r_+ < \infty$ with $r_j \in [r_-, r_+]$ throughout. Fix $s_1 := \min\{\frac{1}{2}, \frac{1}{2r_+}\}$ and $s_0 := s_1/2$. For $(s, y) \in [s_0, s_1] \times \bar{Y}_i$, the preimage $T_\xi^{-1}(s, y)$ has coordinates $a_1 = (\phi')^{-1}(s)$ and $a_j = (\phi')^{-1}(sr_j)$. These are well defined because $sr_j \le \frac{1}{2}$, and they lie in the compact set

$$\mathcal{K} := \big[(\phi')^{-1}(s_1), (\phi')^{-1}(s_0)\big] \times \prod_{j\neq 1}\big[(\phi')^{-1}(s_1 r_+), (\phi')^{-1}(s_0 r_-)\big] \subset (0,\infty)^{d_h},$$

independently of $\xi \in \Xi_i$. On the compact set $[s_0, s_1] \times \bar{Y}_i \times \Xi_i$, the Jacobian determinant of T_ξ^{-1} is continuous and non-vanishing, hence bounded in modulus below by some $j_- > 0$. Therefore, for every Borel $A \subseteq B_i$, Lemma D.4 and the change of variables formula give

$$\mathbb{P}\big([D(a_1)\bar{\xi}] \in A \mid h_0 = h\big) \;\ge\; \int_{T_{\bar{\xi}}^{-1}([s_0,s_1]\times\chi(A))} q(h,a)\,\mathrm{d}a \;\ge\; q_{\mathcal{K}}\,(s_1 - s_0)\,j_-\,\mathrm{Leb}\big(\chi(A)\big),$$

uniformly in $h \in \mathcal{X}$ and $\xi \in \Xi_i$. Since χ is a diffeomorphism whose Jacobian is bounded away from zero and infinity on the compact set B_i, there is $c_\chi > 0$ with $\mathrm{Leb}(\chi(A)) \ge c_\chi \mathrm{vol}(A)$, and $\rho_i := q_{\mathcal{K}}(s_1 - s_0)j_- c_\chi$ has the required property. □

Proposition D.8 (Doeblin condition on $\mathcal{W}$)**.** *There exist $\delta \in (0, 1]$ and a probability measure ϱ on $\mathcal{W}$ with $\bar{Q}^5(w, \cdot) \ge \delta\varrho(\cdot)$ for every $w \in \mathcal{W}$. Consequently $\bar{Q}$ is uniformly ergodic in the sense of Theorem B.8: it is positive Harris and aperiodic, has a unique invariant probability measure ν, and satisfies* $\sup_w \|\bar{Q}^m(w, \cdot) - \nu\|_{\mathrm{TV}} \le 2(1 - \delta)^{\lfloor m/5\rfloor}$.

Proof. Let η_i and ρ_i be as in Lemma D.7, and put $\rho := \min_i \rho_i$ and $\eta := \min\{\min_i \eta_i, \pi/9\}$, so that Lemmas D.6 and D.7 both apply with this common η. Let c_0 and c_1 be as in equation 26 and equation 35. Fix an arbitrary $w_0 = (h_0, \bar{v}_0) \in \mathcal{W}$ and, using equation 27, choose i with $|(W_h\bar{v}_0)_i| \ge c_0$. By the Markov property, the conditional probabilities of successive steps below multiply.

Steps 1–3: all directions reach a common neighbourhood of $[e_1]$. Lemma D.6 with $c = c_0$ gives $\mathrm{d}([\bar{v}_1], [e_i]) \le \eta$ with conditional probability at least $p_A(\eta, c_0)$. On this event, Lemma D.7 gives $[\bar{v}_2] \in B_i$ with conditional probability at least $\rho\,\mathrm{vol}(B_i)$. On that event, equation 35 allows us to apply Lemma D.6 once more, now with index 1 and $c = c_1$, which gives $\mathrm{d}([\bar{v}_3], [e_1]) \le \eta$ with conditional probability at least $p_A(\eta, c_1)$. Multiplying, and minimising over the finitely many possible i,

$$\mathbb{P}\big(\mathrm{d}([\bar{v}_3], [e_1]) \le \eta \mid w_0\big) \;\ge\; p_C \;:=\; p_A(\eta, c_0) \cdot \rho \min_i \mathrm{vol}(B_i) \cdot p_A(\eta, c_1) \;>\; 0,$$

uniformly in w_0. Call this event $E_3 \in \mathcal{F}_3$.

Steps 4–5: a joint density in state and direction. A single step cannot produce such a density: given w_{n-1}, the state $w_n = \Theta(w_{n-1}, a_n)$ is a function of the d_h-dimensional pre-activation a_n alone, whereas $\mathcal{W}$ has dimension $2d_h - 1$. Two steps suffice, because the direction and the pre-activation are then randomised by independent inputs. Let C be the closed unit ball of $\mathbb{R}^{d_h}$. On E_3, Lemma D.7 with $i = 1$ shows that, conditionally on $\mathcal{F}_3$, the law of $[\bar{v}_4]$ dominates $\rho\,\mathrm{vol}$ on B_1. Conditionally on $\mathcal{F}_4$, the law of a_5 has density $q(h_4, \cdot)$, and $q(h_4, \cdot) \ge q_C$ on C *uniformly in* h_4 by Lemma D.4. These two bounds combine into a bound on the joint law of $([\bar{v}_4], a_5)$. Let $G \subseteq B_1 \times C$ be Borel and let $G_v := \{a : (v, a) \in G\}$ be its section at v. Conditioning first on $\mathcal{F}_4$ and then on $\mathcal{F}_3$, we obtain on E_3

$$\begin{aligned}\mathbb{P}\big(([\bar{v}_4], a_5) \in G \mid \mathcal{F}_3\big) &= \mathbb{E}\Big[\int_{G_{[\bar{v}_4]}} q(h_4, a)\,\mathrm{d}a \;\Big|\; \mathcal{F}_3\Big] \\ &\ge\; q_C\,\mathbb{E}\big[\mathrm{Leb}(G_{[\bar{v}_4]}) \mid \mathcal{F}_3\big] \;\ge\; q_C\,\rho \int_{B_1} \mathrm{Leb}(G_v)\,\mathrm{vol}(\mathrm{d}v).\end{aligned}$$

The last inequality extends the set-wise bound of Lemma D.7 to the non-negative measurable function $v \mapsto \mathrm{Leb}(G_v)\mathbf{1}_{B_1}(v)$, by approximation with simple functions. By Tonelli's theorem the right-hand side equals $\rho\, q_C\, (\mathrm{vol}|_{B_1} \otimes \mathrm{Leb}|_C)(G)$. The conditional law of $([\bar{v}_4], a_5)$ therefore dominates $\rho\, q_C\, (\mathrm{vol}|_{B_1}) \otimes (\mathrm{Leb}|_C)$. Since $w_5 = \Theta((h_4, \bar{v}_4), a_5)$ and Θ does not depend on h_4, the conditional law of w_5 dominates the image of this product measure under Θ. Combining with Steps 1–3 gives $\bar{Q}^5(w_0, \cdot) \geq \delta\varrho$ with

$$\varrho := \frac{\Theta_*\big((\mathrm{vol}|_{B_1}) \otimes (\mathrm{Leb}|_C)\big)}{\mathrm{vol}(B_1)\mathrm{Leb}(C)}, \qquad \delta := p_C\, \rho\, q_C\, \mathrm{vol}(B_1)\, \mathrm{Leb}(C),$$

uniformly in w_0; necessarily $\delta \leq 1$, since both $\bar{Q}^5(w_0, \cdot)$ and ϱ are probability measures. The remaining assertions follow from Theorem B.8. □

Corollary D.9 (The enlarged chain)**.** *$\mathcal{Y}$ is small for P: with δ, ϱ, C and q_C as above, $P^6\big((w,a),\cdot\big) \geq \delta\, q_C\, \varrho \otimes (\mathrm{Leb}|_C)$ for every $(w,a) \in \mathcal{Y}$. Consequently P is uniformly ergodic, and V-uniformly ergodic for the drift function equation 34, with a unique invariant probability measure π satisfying $\pi(V) \leq C_0 < \infty$. The $\mathcal{W}$-marginal of π is ν, and $\pi(f) = \nu(F)$.*

Proof. Given $\Phi_n = ((h, \bar{v}), a)$, the state $w_n = \Theta((h, \bar{v}), a)$ is determined, and the law of w_{n+5} given w_n is $\bar{Q}^5(w_n, \cdot) \geq \delta\varrho$ by Proposition D.8. Conditionally on w_{n+5}, the law of a_{n+6} has density at least q_C on C, uniformly in the hidden coordinate of w_{n+5} (Lemma D.4). The conditioning argument of Steps 4–5 above then gives the stated minorisation of P^6, so Theorem B.8 applies to P, and Lemma D.5 together with Corollary B.10 upgrades this to V-uniform ergodicity with $\pi(V) \leq C_0$.

For the marginal, suppose $\Phi_n \sim \pi$ and let π_1 be the law of w_{n-1}. Stationarity implies that w_n, the first coordinate of Φ_{n+1}, also has law π_1, and its conditional law given w_{n-1} is $\bar{Q}(w_{n-1}, \cdot)$; hence $\pi_1\bar{Q} = \pi_1$, and uniqueness in Proposition D.8 gives $\pi_1 = \nu$. Finally, under π the conditional law of a_1 given w_0 is $q(h_0, \cdot)$, so $\pi(f) = \mathbb{E}_\pi[f(w_0, a_1)] = \mathbb{E}_\pi[F(w_0)] = \nu(F)$ by conditioning; this is legitimate since $\pi(|f|) \leq \pi(V)^{1/2} < \infty$. □

D.5 Identification of λ_1 and the central limit theorem

Proposition D.10 (The drift of the additive functional is the top exponent)**.** *$\nu(F) = \pi(f) = \lambda_1$. In particular λ_1 is finite, with $|\lambda_1| \leq \sup_{\mathcal{W}} |F|$.*

Proof. Fix $h_0 \in \mathcal{X}$ and $v \neq 0$, and start the chain equation 31 from $w_0 = (h_0, [v])$. By Corollary D.9 the chain is positive Harris with $\pi(|f|) < \infty$, so Theorem B.11(i) and equation 32 give, almost surely,

$$\frac{1}{m} \log \frac{\|J_{0\to m} v\|}{\|v\|} \;=\; \frac{1}{m} \sum_{n \leq m} f(\Phi_n) \;\longrightarrow\; \pi(f). \tag{36}$$

Apply this with $v = e_i$ for each $i = 1, \ldots, d_h$. For any matrix M, $\max_i \|Me_i\| \leq \|M\| \leq \|M\|_F \leq \sqrt{d_h} \max_i \|Me_i\|$, so

$$\max_i \log \|J_{0\to m} e_i\| \;\leq\; \log \|J_{0\to m}\| \;\leq\; \tfrac{1}{2} \log d_h + \max_i \log \|J_{0\to m} e_i\|.$$

Dividing by m, the two outer terms converge almost surely to $\pi(f)$ by equation 36, while the middle term converges almost surely to λ_1 by Theorem 3.1, which holds under the present hypotheses by Lemma D.1. Hence $\lambda_1 = \pi(f) = \nu(F)$, which is finite because F is bounded (Lemma D.5). □

Remark D.11. In the classical theory of products of random matrices, identifying the growth rate of a fixed vector with that of the operator norm requires a regularity property of the stationary measure on projective space, typically that it charges no projective hyperplane (Bougerol & Lacroix, 1985). No such property is needed here. The bound $\|M\| \leq \sqrt{d_h} \max_i \|Me_i\|$ reduces the operator norm to finitely many fixed directions, and each of them obeys the law of large numbers *from every initial state*, because Proposition D.8 makes the projective chain positive Harris. An ergodic theorem for the stationary chain would give this only for ν-almost every initial direction. This is also the only point at which the operator norm enters: Theorem 4.1, Proposition 4.2 and Corollary 4.3 all concern a fixed transported vector v.

Lemma D.12 (Bounded corrector and martingale decomposition)**.** *The series $\psi := \sum_{n\geq 0}\big(\bar{Q}^n F - \nu(F)\big)$ converges uniformly on $\mathcal{W}$ and defines a bounded measurable solution of the Poisson equation $\psi - \bar{Q}\psi = F - \nu(F)$ for the projective chain. Moreover,*

$$\hat{f}\big((h,\bar{v}),a\big) := f\big((h,\bar{v}),a\big) - \lambda_1 + \psi\big(\Theta((h,\bar{v}),a)\big) \tag{37}$$

solves Poisson's equation $\hat{f} - P\hat{f} = f - \pi(f)$ for the enlarged chain, with $P\hat{f}(w,a) = \psi(\Theta(w,a))$ and $\pi(\hat{f}^2) < \infty$. With the convention $P\hat{f}(\Phi_0) := \psi(w_0)$, the increments

$$\Delta_n \;:=\; \hat{f}(\Phi_n) - P\hat{f}(\Phi_{n-1}) \;=\; f(w_{n-1},a_n) - \lambda_1 + \psi(w_n) - \psi(w_{n-1}) \tag{38}$$

form a square-integrable martingale difference sequence with respect to $\{\mathcal{F}_n\}$, and

$$\sum_{n=1}^{m} f(w_{n-1},a_n) \;=\; \lambda_1 m \;+\; \sum_{n=1}^{m}\Delta_n \;+\; \psi(w_0) - \psi(w_m). \tag{39}$$

Proof. F is bounded (Lemma D.5), and Proposition D.8 gives $|\bar{Q}^n F(w) - \nu(F)| \leq 2\|F\|_\infty(1-\delta)^{\lfloor n/5\rfloor}$ uniformly in w. The series therefore converges uniformly, $\|\psi\|_\infty < \infty$, and the Poisson identity follows by applying $\bar{Q}$ term by term and telescoping.

For the enlarged chain, let $w' = \Theta((h,\bar{v}),a)$ be the state reached from $\Phi_n = ((h,\bar{v}),a)$. Then $P\hat{f}(\Phi_n) = \mathbb{E}[\hat{f}(\Phi_{n+1}) \mid \Phi_n] = F(w') - \lambda_1 + \bar{Q}\psi(w')$, and substituting $\bar{Q}\psi = \psi - F + \nu(F)$ with $\nu(F) = \lambda_1$ (Proposition D.10) gives $P\hat{f}(\Phi_n) = \psi(w')$. Subtracting, $\hat{f} - P\hat{f} = f - \lambda_1 = f - \pi(f)$. Square integrability follows from

$$\pi(\hat{f}^2) \leq 2\pi\big((f-\lambda_1)^2\big) + 2\|\psi\|_\infty^2 \leq 4\pi(V) + 4\lambda_1^2 + 2\|\psi\|_\infty^2 < \infty,$$

using $f^2 \leq V$ and $\pi(V) < \infty$ (Lemma D.5, Corollary D.9). The same bound, with $\sup_{\mathcal{Y}} PV < \infty$ in place of $\pi(V) < \infty$, shows that each Δ_n is square integrable from any initial state. The explicit form equation 38 is equation 37 minus $P\hat{f}(\Phi_{n-1}) = \psi(w_{n-1})$, and the martingale property is

$$\mathbb{E}[\Delta_n \mid \mathcal{F}_{n-1}] = F(w_{n-1}) - \lambda_1 + \bar{Q}\psi(w_{n-1}) - \psi(w_{n-1}) = 0.$$

Summing equation 38 over n and telescoping the ψ terms gives equation 39. □

Remark D.13 (Why the splitting matters). The summand f is unbounded below, so the generic construction of Meyn & Tweedie (2009, Thm. 17.4.2), which bounds the Poisson solution by a multiple of $V+1$, would produce an unbounded corrector. Lemma D.12 instead exploits the structure of the enlarged chain. The fresh pre-activation enters only the current summand, so $P^n f = (\bar{Q}^{n-1}F)\circ\Theta$ is bounded for every $n \geq 1$, and the Poisson solution splits as $\hat{f} = (f-\lambda_1) + \psi\circ\Theta$ with ψ *bounded*. This boundedness is essential: it is exactly what the proof of Proposition 4.2 contradicts (Appendix D.6), and without it that proof does not go through.

Proposition D.14 (CLT and LIL along a fixed direction)**.** *Let $\sigma^2 := \gamma_f^2$ be the constant equation 15 of Theorem B.11 for $g = f$. Then $\sigma^2 \in [0,\infty)$ is given by equation 33 and satisfies $\sigma^2 = \pi\big(\hat{f}^2 - (P\hat{f})^2\big) = \mathbb{E}_\pi[\Delta_1^2]$. Moreover, for every $h_0 \in \mathcal{X}$ and every $v \neq 0$, with $S_m := \log(\|J_{0\to m}v\|/\|v\|)$,*

$$\frac{S_m - \lambda_1 m}{\sqrt{m}} \;\xrightarrow{\;d\;}\; \mathcal{N}(0,\sigma^2),$$

and, if $\sigma > 0$, almost surely

$$\limsup_{m\to\infty} \frac{S_m - \lambda_1 m}{\sqrt{2m\log\log m}} = \sigma, \qquad \liminf_{m\to\infty} \frac{S_m - \lambda_1 m}{\sqrt{2m\log\log m}} = -\sigma. \tag{40}$$

Proof. By Corollary D.9 the chain Φ is V-uniformly ergodic, and $f^2 \leq V$ by Lemma D.5, so Theorem B.11(ii) applies to $g = f$. The constant γ_f^2 of equation 15 is therefore well defined and finite, and since $\pi(f) = \lambda_1$ and $f(\Phi_n) = f_n$, it coincides with the right-hand side of equation 33. Lemma D.12 provides a solution of Poisson's equation with $\pi(\hat{f}^2) < \infty$, so Theorem B.12 gives $\gamma_f^2 = \pi(\hat{f}^2 - (P\hat{f})^2) = \mathbb{E}_\pi[\Delta_1^2]$.

If $\gamma_f^2 > 0$, Theorem B.11(iv) gives the central limit theorem and the law of the iterated logarithm from every initial condition. If $\gamma_f^2 = 0$, Theorem B.11(iii) gives $m^{-1/2}(S_m - \lambda_1 m) \to 0$ almost surely, which is convergence in distribution to the degenerate law $\mathcal{N}(0,0)$. In both cases the conclusion holds from every deterministic initial state $\Phi_1 = (w_0, a_1)$, and hence, by conditioning and dominated convergence, from any initial law; in particular from $w_0 = (h_0, [v])$ with $a_1 \sim q(h_0, \cdot)$. □

Proof of Theorem 4.1. Finiteness of λ_1 is Proposition D.10, and the case $t = 0$ is Proposition D.14. For a general time origin t, the vector $J_{t\to t+m}v$ is obtained by running equation 28 from the state $w_t = (h_t, [v])$, whose law is a probability measure on $\mathcal{W}$ determined by h_0 and t. Proposition D.14 holds from every deterministic initial state, and convergence in distribution is preserved under mixing over the initial law: for bounded continuous φ, apply dominated convergence to $w \mapsto \mathbb{E}_w[\varphi((S_m - \lambda_1 m)/\sqrt{m})]$. The same holds for the almost sure statement equation 40, since an event of probability one from every initial state has probability one under any mixture of initial states. Finally, $\sigma^2 = \gamma_f^2$ depends only on π and f, and hence only on $(W_h, W_u, b_h, \phi, \Gamma)$. □

D.6 Proof of Proposition 4.2

Proof. Suppose $\sigma^2 = \mathbb{E}_\pi[\Delta_1^2] = 0$. By Proposition D.14, $\Delta_1 = 0$ holds π-a.s., so by equation 38 the one-step log-stretching satisfies Poisson's equation:

$$\log \left\| D(a) W_h \bar{v}_0 \right\| \; + \; \psi\Big(\phi(a), \big[D(a) W_h \bar{v}_0\big]\Big) \; = \; \lambda_1 + \psi(h_0, \bar{v}_0) \tag{41}$$

for ν-a.e. $w_0 = (h_0, \bar{v}_0)$ and Lebesgue-a.e. $a \in \mathbb{R}^{d_h}$; the passage from $q(h_0, \cdot)$-a.e. to Lebesgue-a.e. a uses $q(h_0, \cdot) > 0$ everywhere (Lemma D.4). The right-hand side of equation 41 does not depend on a, and ψ is bounded (Lemma D.12). The idea of the proof is to change a so that the first term on the left moves by an arbitrarily large amount while the arguments of ψ change only in the hidden-state coordinate.

1. Coordinatewise dilation. Fix $\varepsilon \in (0,1)$ and define $S_\varepsilon : \mathbb{R}^{d_h} \to \mathbb{R}^{d_h}$ coordinatewise by

$$(S_\varepsilon a)_i := \operatorname{sign}(a_i)\, (\phi')^{-1}\big(\varepsilon \phi'(a_i)\big),$$

with $(\phi')^{-1}$ as in Assumption 5. Since $\varepsilon\phi'(a_i) \in (0, \varepsilon]$, each coordinate of $S_\varepsilon a$ is non-zero and has the sign of a_i, so S_ε maps each open orthant into itself. On each open orthant, S_ε is a C^1 diffeomorphism onto its image with diagonal Jacobian entries

$$\frac{\varepsilon\, |\phi''(a_i)|}{\big|\phi''\big((\phi')^{-1}(\varepsilon\phi'(a_i))\big)\big|} \; > \; 0,$$

which are finite and non-zero because $a_i \neq 0$ and $(\phi')^{-1}(\varepsilon\phi'(a_i)) > 0$. The coordinate hyperplanes are Lebesgue-null, so S_ε and S_ε^{-1} map Lebesgue-null sets to Lebesgue-null sets.

2. Norm scaling and projective invariance. Since ϕ' is even, $\phi'((S_\varepsilon a)_i) = \varepsilon\phi'(a_i)$ for every i, that is, $D(S_\varepsilon a) = \varepsilon D(a)$. Consequently

$$\begin{aligned} \left\| D(S_\varepsilon a) W_h \bar{v}_0 \right\| &= \varepsilon \left\| D(a) W_h \bar{v}_0 \right\|, \\ \big[D(S_\varepsilon a) W_h \bar{v}_0\big] &= \big[D(a) W_h \bar{v}_0\big] =: \bar{v}_1 \in \mathbb{P}^{d_h - 1}. \end{aligned}$$

Replacing a by $S_\varepsilon a$ in equation 41 therefore shifts the first term by $\log\varepsilon$, while the direction $\bar{v}_1$ in the argument of ψ is unchanged; only the hidden coordinate moves, from $\phi(a)$ to $\phi(S_\varepsilon a)$.

3. Contradiction. Fix w_0 outside the ν-null exceptional set, and let $N \subset \mathbb{R}^{d_h}$ be the Lebesgue-null set of a at which equation 41 fails. By Step 1, $N \cup S_\varepsilon^{-1}(N)$ is Lebesgue-null, so there is $a \in \mathbb{R}^{d_h} \setminus \big(N \cup S_\varepsilon^{-1}(N)\big)$ at which equation 41 holds both at a and at $S_\varepsilon a$. Subtracting the identity at $S_\varepsilon a$ from the identity at a cancels the right-hand side and, by Step 2, gives

$$\log\varepsilon \; = \; \psi\big(\phi(a), \bar{v}_1\big) - \psi\big(\phi(S_\varepsilon a), \bar{v}_1\big), \qquad \text{hence} \qquad |\log\varepsilon| \; \le \; 2\|\psi\|_\infty.$$

Since $\varepsilon \in (0,1)$ was arbitrary and $|\log\varepsilon| \to \infty$ as $\varepsilon \to 0^+$, this contradicts the boundedness of ψ. Hence $\sigma^2 > 0$. □

The proof shows what $\sigma = 0$ would require: the one-step log-stretching would have to equal λ_1 up to the coboundary of a *bounded* function ψ (Remark D.13). Equivalently, after lengths at each state are rescaled by the bounded factor $e^{\psi(h,[v])}$, every one-step Jacobian would stretch every direction by exactly e^{λ_1}. A pointwise two-sided bound $c \le \sigma_{d_h}(A) \le \sigma_1(A) \le C$, of the kind imposed by the architectural families of Section 1.1, does not provide this: it bounds the increments of $\log \|J_{0\to m} v\|$ but not their fluctuations, and a bounded i.i.d. sequence still obeys a non-degenerate central limit theorem. An exactly orthogonal linear recurrence ($\phi' \equiv 1$) does satisfy the condition, with $\psi \equiv 0$ and $\lambda_1 = 0$; a saturating ϕ rules it out.

Among recurrences with $\lambda_1 = 0$, those that are isometric in the above sense form a strict subclass of those obeying a pointwise two-sided bound, which in turn form a strict subclass of all recurrences with $\lambda_1 = 0$; $\sigma = 0$ is attainable only in the first class. Proposition 4.2 shows that no vanilla RNN with a saturating activation belongs to it, for any W_h, orthogonal W_h included.

D.7 Proof of Corollary 4.3

Proof. Let $\Phi_{\mathcal{N}}$ denote the standard normal distribution function, and recall that $\sigma > 0$ (Proposition 4.2). By Theorem 4.1, $G_m := (\log R_m - \lambda_1 m)/(\sigma\sqrt{m})$ converges in distribution to $\mathcal{N}(0,1)$. Because $\Phi_{\mathcal{N}}$ is continuous, this convergence is uniform (Pólya's theorem): $\sup_{x\in\mathbb{R}} |\mathbb{P}(G_m \le x) - \Phi_{\mathcal{N}}(x)| \to 0$, and likewise with $<$ in place of $\le$. Consequently, if $x_m \to c \in [-\infty, \infty]$, possibly with x_m depending on further parameters, then

$$\mathbb{P}(G_m \le x_m) \;\to\; \Phi_{\mathcal{N}}(c), \qquad \Phi_{\mathcal{N}}(-\infty) := 0, \quad \Phi_{\mathcal{N}}(\infty) := 1, \tag{42}$$

and likewise for $\mathbb{P}(G_m < x_m)$. Both parts of the corollary reduce to equation 42.

(i) Let $\lambda_1 = 0$, so that $G_m = \log R_m/(\sigma\sqrt{m})$, and let p be a polynomial with positive leading coefficient. Then $p(m) > 0$ for large m and $\log p(m) = O(\log m) = o(\sqrt{m})$, so $x_m := \log p(m)/(\sigma\sqrt{m}) \to 0$. By equation 42,

$$\mathbb{P}\big(R_m > p(m)\big) = 1 - \mathbb{P}(G_m \le x_m) \;\to\; 1 - \Phi_{\mathcal{N}}(0) = \tfrac{1}{2},$$
$$\mathbb{P}\big(R_m < 1/p(m)\big) = \mathbb{P}(G_m < -x_m) \;\to\; \Phi_{\mathcal{N}}(0) = \tfrac{1}{2}.$$

(ii) Write $z := z_{1-\alpha/2}$, so that $2\Phi_{\mathcal{N}}(z) - 1 = 1 - \alpha$. The left-hand side of equation 11 is continuous and strictly increasing in $m \ge 0$, vanishes at $m = 0$, and diverges as $m \to \infty$; hence equation 11 has a unique positive root $m^* = m^*(\kappa)$, and $m^*(\kappa) \to \infty$ as $\kappa \to \infty$. For $m \ge 1$,

$$\mathbb{P}\big(R_m \in [\kappa^{-1}, \kappa]\big) = \mathbb{P}\big(G_m \le c_m^+\big) - \mathbb{P}\big(G_m < c_m^-\big), \qquad c_m^\pm := \frac{\pm \log\kappa - \lambda_1 m}{\sigma\sqrt{m}}.$$

We let $\kappa \to \infty$ and $m = m_\kappa \to \infty$ with $m_\kappa/m^*(\kappa) \to r \in (0,\infty)$, and apply equation 42 to $c_m^\pm$.

If $\lambda_1 = 0$, then $m^* = (\log\kappa/(z\sigma))^2$ and $c_m^\pm = \pm z\sqrt{m^*/m} \to \pm z/\sqrt{r}$, so the probability converges to $2\Phi_{\mathcal{N}}(z/\sqrt{r}) - 1$. This limit exceeds $1-\alpha$ for $r < 1$, equals $1-\alpha$ for $r = 1$, and falls below $1-\alpha$ for $r > 1$.

If $\lambda_1 > 0$ (the case $\lambda_1 < 0$ is identical with the two tails exchanged), then $c_m^- \le -\lambda_1\sqrt{m}/\sigma \to -\infty$, so only the upper tail matters. Using $\log\kappa = \lambda_1 m^* + z\sigma\sqrt{m^*}$ from equation 11,

$$c_m^+ \;=\; \frac{\lambda_1(m^* - m)}{\sigma\sqrt{m}} \;+\; z\sqrt{\frac{m^*}{m}}.$$

For $r < 1$ the first term tends to $+\infty$, and for $r > 1$ to $-\infty$, so the probability tends to 1 and 0 respectively. At $m = \lfloor m^* \rfloor$ the first term vanishes in the limit, $c_m^+ \to z$, and the probability tends to $\Phi_{\mathcal{N}}(z) = 1 - \alpha/2 \ge 1 - \alpha$.

In both cases, horizons shorter than $m^*(\kappa)$ by a fixed factor keep R_m within $[\kappa^{-1}, \kappa]$ with limiting probability above $1-\alpha$, and horizons longer by a fixed factor do not. In this sense $m^*(\kappa)$ is the usable horizon to leading order. □

Two comments on part (ii). When $\lambda_1 \ne 0$ only one tail is active at the horizon, so replacing $z_{1-\alpha/2}$ by $z_{1-\alpha}$ in equation 11 makes the limiting probability at $m = \lfloor m^* \rfloor$ exactly $1-\alpha$. Both choices give

the same leading order $m^* \sim \log \kappa / |\lambda_1|$, with σ entering through the correction of order $\sqrt{\log \kappa}$; we state equation 11 with the two-sided quantile because it then covers every value of λ_1 with a single formula, conservatively when $\lambda_1 \neq 0$.

The same arguments give two further consequences at $\lambda_1 = 0$. First, for $p \in (0, 1)$ the p-quantile of R_m is $\exp\big(\sigma\sqrt{m}\,(z_p + o(1))\big)$: convergence in distribution to a limit with continuous, strictly increasing distribution function implies convergence of quantiles, and $x \mapsto e^{\sigma\sqrt{m}\,x}$ is increasing. Second, the law of the iterated logarithm equation 40 shows that, for every $\eta > 0$, almost every trajectory has $R_m > e^{(1-\eta)\sigma\sqrt{2m \log\log m}}$ and $R_m < e^{-(1-\eta)\sigma\sqrt{2m \log\log m}}$ each for infinitely many m: a single transported gradient both explodes and vanishes, infinitely often. Like Theorem 4.1, all of these statements concern one fixed transported vector v.

D.8 The batch-means estimator

Split the penalty window of length m into B consecutive blocks of length $\ell = m/B$, let $\bar{f}^{(b)}$ be the mean of the per-step log-stretching over block b, and let $\bar{f}$ be the mean over the whole window, which is the finite-time exponent. The estimator of Section **??** is

$$\widehat{\sigma}^2 \;=\; \frac{\ell}{B-1} \sum_{b=1}^{B} \big(\bar{f}^{(b)} - \bar{f}\big)^2. \tag{43}$$

This is the standard batch-means estimator of the asymptotic variance equation 33 of the additive functional equation 32 of the enlarged chain.

Consistency. Strong consistency of batch means requires $\ell \to \infty$ and $B \to \infty$, geometric ergodicity of the chain, and a moment condition on the summand whose strength depends on how fast ℓ grows (Flegal & Jones, 2010). Geometric ergodicity, in fact uniform ergodicity, is supplied by Corollary D.9. For the moments, Lemma D.5 gives $|f| \le C_1 + L(a)$, and for $\phi = \tanh$ the bound $|\log \phi'(x)| \le C_\phi(1 + |x|)$ yields $|f(\Phi_n)| \le C(1 + \|u_n\|)$ for a constant C: the summand has as many moments as the input. With block lengths $\ell = \lfloor m^\theta \rfloor$, higher moments admit smaller exponents $\theta \in (0, 1)$, and a moment of every order admits every $\theta \in (0, 1)$ (Flegal & Jones, 2010). For Gaussian or sub-Gaussian inputs, which have moments of every order, the usual choice $\ell = \lfloor \sqrt{m} \rfloor$ is therefore admissible; under the $(2 + \epsilon)$-moment of Assumption 6 alone, consistency is guaranteed only for longer blocks.

Bias. At fixed m the estimator is biased downwards. Let $R_n := \mathrm{Cov}_\pi(f_1, f_{1+n})$, which decays geometrically by V-uniform ergodicity (Meyn & Tweedie, 2009, Thm. 16.1.5). Under the stationary law, the standard expansion of the batch-means bias is

$$\mathbb{E}\widehat{\sigma}^2 \;=\; \sigma^2 \;-\; \frac{\Gamma_1}{\ell} \;+\; o\Big(\frac{1}{\ell}\Big), \qquad \Gamma_1 := 2\sum_{n \ge 1} n\, R_n,$$

with the series converging absolutely, while the variance of $\widehat{\sigma}^2$ is of order $\ell/m = 1/B$. The bias is negative exactly when $\Gamma_1 > 0$, that is, when the short-lag correlations of the log-stretching are predominantly positive, which is what one expects and what we observe. In practice, short blocks therefore under-estimate σ^2 and under-regularise.

E Direct numerical tests of the theory

Experiments in E.1 and E.2 test Theorem 3.1 and Theorem 4.1 respectively. Neither requires training: each concerns the map $\vartheta \mapsto (\lambda_1, \sigma)$ at a fixed ϑ, so a randomly initialized network works as a test object, and keeps these tests independent of the training results of Section 5. Unless stated otherwise the network is an untrained vanilla RNN with $d_h = 128$ and $W_{ij} \sim \mathcal{N}(0, 1/d_h)$, all measurement in float64 with a 200-step frame burn-in, driven by the psMNIST input stream. Every panel below costs minutes on one CPU.

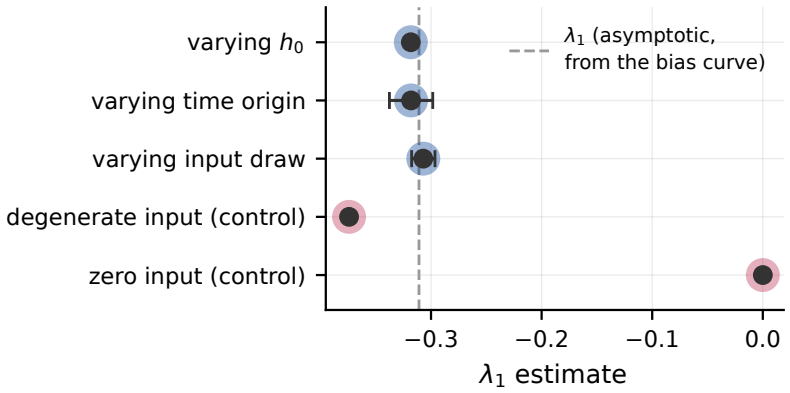

Figure 7: **Theorem 3.1, tested against the input.** Blue: estimates of λ_1 at $m = 4000$ over 16 replicates each, varying only what the theorem says should *not* matter — the initial state h_0, the time origin along one long trajectory, and the input realization; dashed line: the asymptotic λ_1 measured at the longest horizon ($m = 2 \times 10^4$). The three agree to within 0.011. Red: two changes to what the theorem says *should* matter, the input law itself, both violating Assumption 1. Confining the input to a one-dimensional subspace shifts the exponent to -0.374 ($+22\%$); removing the input drive entirely moves it to ≈ 0, two orders of magnitude from the driven value.

E.1 Testing Theorem 3.1 against the input

Theorem 3.1 asserts that λ_1 depends on the input only through its law, not on the initial state, the time origin, or the particular realization drawn. Figure 7 tests this directly: estimates of λ_1 at $m = 4000$ over 16 replicates each, varying h_0, the time origin along one long trajectory, and the input realization, give -0.318 (s.d. 3×10^{-4}), -0.318 (s.d. 2.0×10^{-2}) and -0.307 (s.d. 1.1×10^{-2}) respectively — agreement to within 0.011, against an asymptotic reference measured at $m = 2 \times 10^4$. The spread shrinks with horizon at a rate consistent with $m^{-1/2}$: over $m = 500, \ldots, 4000$ the product $\text{s.d.} \times \sqrt{m}$ stays in $[0.62, 0.94]$ without trend, though 16 trajectories do not pin it down more tightly than that.

The theorem is a statement about the law of the input, not about any one realization of it, so the sharper test is to change the law itself. Two such changes, both violating Assumption 1, do move λ_1: confining the input to a one-dimensional subspace shifts the exponent to -0.374 ($+22\%$), and removing the input drive entirely moves it to ≈ 0, two orders of magnitude from the driven value. The degenerate-input-law case behaves as one would expect once Assumption 1 fails: a genuine, reliable shift in λ_1. The zero-input case is more interesting: its spread *across* h_0 stays at 7×10^{-5}, so read as a test of h_0-dependence alone at fixed (degenerate) input law — which is how a multistability argument would motivate it — it looks like a pass rather than a failure. Simulating the autonomous map $h \mapsto \tanh(W_h h)$ directly confirms that the zero-input system is multistable (different h_0 reach different fixed points, with $\rho(\text{diag}(\phi'(a^*))W_h)$ between 0.991 and 0.996), but those fixed points are near-marginal and similar enough to each other that λ_1 barely distinguishes them. The lesson for testing Theorem 3.1 this way: the informative signal is the shift in λ_1 when the input law itself changes, not the spread across initial conditions at a fixed law; we lead with the degenerate-input-law case for that reason.

E.2 Testing Theorem 4.1 against an independently measured σ

Figure 2, which appears in the main text, is the experiment attached to Theorem 4.1; this subsection records the details. It is worth being precise about what makes it a test rather than a fit: σ enters from a separate measurement on separate data, so the dashed line has no adjustable scale, and a 10% error would be visible. Repeating the whole procedure on the trained `vanilla` checkpoint gives ratio 0.99 ± 0.03 and slope 0.53 under the i.i.d. law — the CLT holds equally well at the end of training as at initialization. Gaussianity is not the binding issue in either case: across all horizons and both laws the standardized third moment stays below 0.46 and the excess fourth moment below 0.41 in absolute value. A third setting closes the $d_u = 1$ gap directly rather than only via the parameter-free check: an untrained network with $d_u = d_h = 128$ and genuinely i.i.d. Gaussian inputs satisfies every hypothesis of Theorem 4.1 at once. Across three seeds ($\lambda_1 \approx -0.47$) the ratio is 0.98–1.02 and the

slope 0.51–0.53; at a fourth network calibrated to $\lambda_1 = 0.0002$, the one drawn in Figure 2(b), the ratio is 1.00 and the slope 0.48, with skew and excess kurtosis below 0.22 and 0.42 throughout.

The failure under the task law is a real caveat and we state it as one. The psMNIST pixel stream is not i.i.d. in time — all 784 pixels of an image share a digit identity, a correlation that persists across the whole block rather than decaying within it — so Assumption 1 does not hold for it, and Figure 2 shows that this is not a technicality: the diffusive constant is inflated and the growth exponent is pushed from $1/2$ toward $2/3$. What this does *not* do is undermine the comparisons in Section 5, and Table 3 is the reason.

Table 3: **The spectrum under the task law and under the i.i.d. law**, mean over 5 seeds, measured at the end of training. λ_1 is essentially law-independent, as Theorem 3.1 would suggest. σ is not: the i.i.d. law returns a value 10–14% smaller for the vanilla variants and 31% smaller for the orthogonal baseline. The ordering of the presets by σ is identical under the two laws, and so is the size of the effect of the variance penalty: adding it to flossing cuts σ by about two-thirds under either law ($k = 5$: $0.115 \to 0.038$ under the task law, $0.104 \to 0.034$ under the i.i.d. law). Every conclusion in Section 5 is a statement about that ordering, and none of them changes.

	λ_1		σ		
	task law	i.i.d. law	task law	i.i.d. law	ratio
vanilla	-0.0238	-0.0227	0.144	0.128	0.89
floss (k=1)	-0.0208	-0.0199	0.137	0.123	0.90
floss (k=5)	-0.0149	-0.0141	0.115	0.104	0.90
var. only	-0.0067	-0.0063	0.037	0.032	0.86
floss+var. (k=1)	-0.0067	-0.0061	0.037	0.033	0.88
floss+var. (k=5)	-0.0067	-0.0062	0.038	0.034	0.89
orthogonal	-0.0532	-0.0531	0.058	0.040	0.69

E.3 Other recurrent cells at matched drift

Figure 3(b) repeats the measurement of Appendix E.2 on seven untrained cells, $d_h = 128$, $d_u = 1$, i.i.d. Gaussian inputs. The LSTM's state is (h, c), so its QR recursion runs in dimension $2d_h$. The selective model is a one-channel S6-type recurrence, $h' = a_t \odot h + \Delta_t x_t$ with $a_t = \exp(-\Delta_t \lambda)$ and $\Delta_t = \mathrm{softplus}(\cdot)$ of the current input; the LTI model is the same with a constant Δ (S4D-real). Each cell is calibrated to $\lambda_1 = -0.01$ by bisection on one scalar knob, leaving its random weights fixed: the scale of the recurrent matrix (vanilla, orthogonal, LSTM, GRU), the damping γ (antisymmetric), and a common multiplier on the decay rates (both SSMs). Matching at -0.01 rather than 0 is forced: for the diagonal models $\lambda_1 = 0$ is degenerate, as below. σ is measured once from a separate $16{,}000$-step run and the spread at each m from 512 fresh trajectories, exactly as in E4.

Table 4: **Seven recurrences at matched drift** (Figure 3(b)). m^* at $\kappa = 10$; the drift alone caps it at $\log \kappa / |\lambda_1| \approx 230$. "Ratio" is observed over predicted spread, range over $m \in \{250, \ldots, 4000\}$. The last column repeats the measurement under the psMNIST input law.

	knob	λ_1	σ	m^*	slope	ratio	σ (psMNIST)
vanilla	1.761	-0.0080	0.155	42	0.48	0.97–1.04	0.252
orthogonal	1.738	-0.0094	0.138	47	0.49	0.98–1.01	0.237
GRU	3.704	-0.0093	0.080	88	0.51	1.01–1.06	0.078
LSTM	1.918	-0.0099	0.020	180	0.48	0.74–0.83	0.023
antisymmetric	0.374	-0.0099	0.0094	207	0.56	0.69–0.89	0.0055
selective SSM	135.8	-0.0100	0.0020	224	0.51	1.03–1.06	0.0047
LTI SSM	0.200	-0.0100	0	231	—	—	0

Two exact checks. A diagonal recurrence has no mixing between channels, so the top exponent's own series $\log a_{t,\mathrm{top}}$ is observable directly rather than through the QR frame, and its batched-means variance is a second, frame-free estimate of σ^2. For the selective model the two agree to eight digits ($\sigma = 0.0019797262$ against 0.0019797262); for the LTI model $\log a_{\mathrm{top}}$ is constant, so $\sigma = 0$ exactly, and the QR-based spread is below 3×10^{-8} at every m (single-precision rounding). Both require the frame burn-in to exceed the alignment time: the frame converges to the top channel at rate e^{-gt}, with

g the gap between the two largest channel means of $\log a$, and we burn in for $\ln(10^6)/g$ steps (9,255 for LTI, 2,298 for the selective model). With the 1,000-step burn-in used elsewhere, the LTI model reads a spread of 0.5 at $m = 4000$ that is entirely transient.

The degeneracy at $\lambda_1 = 0$. Scaling the decay multiplier of the selective model down by four decades moves (λ_1, σ) from $(-7.4, 1.6) \times 10^{-5}$ to $(-7.4, 1.6) \times 10^{-8}$ and then to $(0, 0)$ together, so the approach to marginal stability is along a ray into the origin, with no point at which $\lambda_1 = 0$ and $\sigma > 0$.

What is less clean. For LSTM and the antisymmetric cell the observed spread is 11–31% below the prediction at every m, while the log–log slopes (0.48, 0.56) stay close to $\frac{1}{2}$. The shape of the law holds; the level is off. Both cells mix slowly (the antisymmetric step is $\varepsilon = 0.05$, and a forget gate near one is a long memory), which biases the batched-means reference at block length 126. Under the psMNIST law the ordering is unchanged, the LTI model stays at $\sigma = 0$, and the slopes inflate to 0.61–0.67, as in Appendix E.2.

E.4 Decorrelating λ_1 from σ, and a probe of regularity

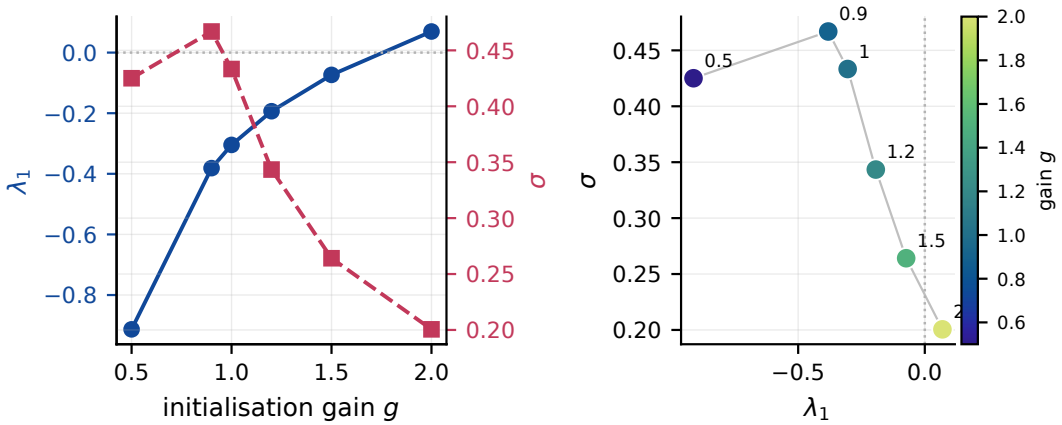

Figure 8: **The initialization gain moves λ_1 across zero without moving σ proportionately.** Untrained vanilla RNNs, $d_h = 128$, at gains $g \in \{0.5, \ldots, 2.0\}$. *Left:* λ_1 (blue) sweeps from -0.91 to $+0.07$, crossing zero near $g \approx 1.75$, while σ (red) varies over a much smaller relative range and non-monotonically, peaking near $g \approx 0.9$. *Right:* the same points in the (λ_1, σ) plane. The two coordinates are not reparametrizations of one another, which is what the collinearity caveat in Section 5.2 needs in order to be addressable.

Figure 8 supplies the untrained half of the gain experiment: it establishes that the (λ_1, σ) plane can be populated off the diagonal that the seven trained runs happen to lie on. It does not by itself show whether accuracy tracks σ rather than λ_1, which needs task accuracy attached to those points and therefore needs the gain grid to be *trained*. Figure 9 supplies that half: all six gains, three seeds each, trained to completion under the same protocol as the main grid — the five stable ones in the panels, the sixth reported beneath them.

Two cautions on how much this carries. The gain grid loosens the coupling between λ_1 and σ; it does not remove it ($r = -0.62$ is still substantial), so the partials are better identified here than in the preset family but not cleanly identified in an absolute sense. And the design varies one scalar at initialization, so it speaks to networks that differ in where they started, not to networks that differ in architecture; the atlas of Section 5.1 is the wider claim and remains the weaker one.

Robustness to the unit of analysis. The correlations above are computed over individual runs. Recomputing them over seed means instead changes nothing material. Writing each family as the triple (coupling $r(\lambda_1, \sigma)$; accuracy on σ given λ_1; accuracy on λ_1 given σ), the preset family gives $(-0.97, +0.22, +0.42)$ per run against $(-0.98, +0.26, +0.45)$ per mean, and the gain family $(-0.59, -0.83, -0.38)$ per run against $(-0.52, -0.90, -0.50)$ per mean: each entry moves by at most 0.07 between units and none changes sign. What the reading does depend on is the *family*. Aligning the presets on a common β^* pushed their coupling to -0.97, so accuracy on σ given

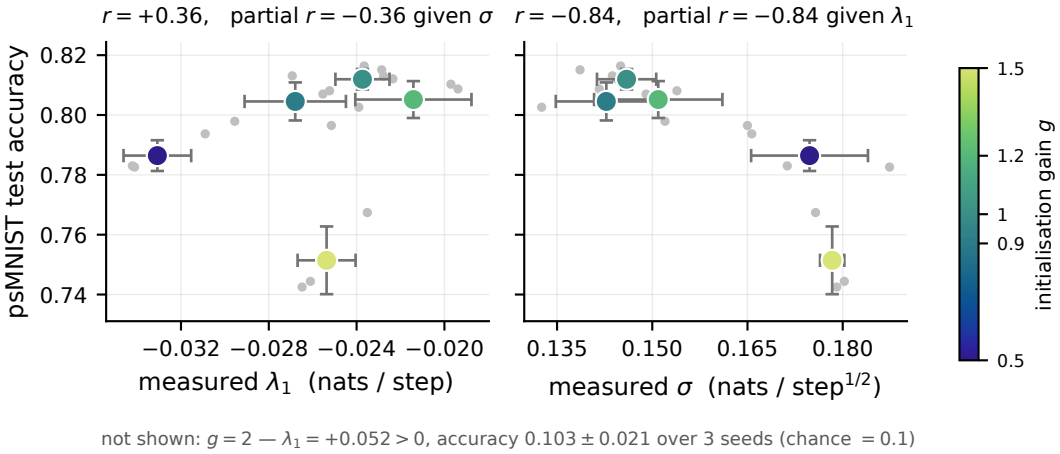

Figure 9: **Initialization gain decorrelates λ_1 from σ, and accuracy follows σ.** Trained vanilla RNNs, $d_h = 128$, at gains $g \in \{0.5, 0.9, 1.0, 1.2, 1.5\}$, three seeds each; large markers are seed means with standard-deviation bars, small gray markers the individual runs. Both panels share the accuracy axis, so the contrast is one of shape and not of zoom. *Left:* against λ_1 the points form a scattered arch, and the marginal correlation ($+0.36$) has the opposite sign to the partial one (-0.36). *Right:* against σ they fall on a monotone decline, and the partial correlation (-0.84) is identical to the marginal one — that is, conditioning on λ_1 removes none of the relationship. Within the six presets of Table 1 these two regressors correlate at -0.94 and the corresponding partials are an uninformative -0.48 and $+0.36$; here they correlate at -0.62, and that is enough to tell them apart. Runs whose measured λ_1 is positive are held out of the panels — the criterion is the measured exponent, not the hyperparameter — and reported beneath them: at $g = 2.0$ the network is above the stability boundary and trains to chance. Left in, that single condition drags the accuracy–λ_1 correlation from $+0.36$ to -0.97 and would support the opposite reading; both numbers are in the sidecar `figG_gain_trained_stats.json`.

λ_1 collapses to $+0.22$ there while holding at -0.83 in the gain family: the quantity the argument rests on is identified in the gain family and simply not identified in the preset family, which is the division of labor Section 5.2 and this appendix assume. A note recorded here in an earlier draft claimed that $n = 18$ was wrong and that the gain family was $n = 15$ (5 gains $\times$ 3 seeds). It is not: the family has 21 runs, of which the three at $g = 2.0$ are held out by measured λ_1, leaving 18. Regenerated from `runs/psmnist/results.csv`, the stable gain family gives $r(\lambda_1, \sigma) = -0.615$, accuracy on σ given λ_1 of -0.837, accuracy on λ_1 given σ of -0.361, and accuracy against m^* of $+0.805$; on all 21 runs $r(\text{acc}, \lambda_1)$ is -0.973. Every figure in this subsection now emits these into `figG_gain_trained_stats.json` alongside the panels.

E.5 Regularity of the spectrum in the parameters

Gradient flossing does something stronger than optimize the exponents: it *differentiates through* them, backpropagating through the QR recursion to obtain $\partial\lambda_i/\partial\vartheta$. For that to be meaningful one needs $\vartheta \mapsto \lambda_i(\vartheta)$ to be continuous, ideally C^1, and the finite-m backpropagated gradient to converge to it. Neither has been established, and the first is genuinely delicate: Lyapunov exponents are discontinuous in the cocycle in general. The optimistic reading is that the pathologies live exactly where our assumptions exclude them — the known counterexamples rely on singular driving measures or on failure of strong irreducibility, whereas here the driving measure has a positive density, every $\mathbf{A}(\omega)$ is invertible, and the base chain is uniformly geometrically ergodic. That is the regime in which Le Page-type Hölder continuity, and in some settings analyticity, is available (Le Page, 1982; Bougerol & Lacroix, 1985; Ruelle, 1979; Duarte & Klein, 2016).

Conjecture E.1 (Regularity of the spectrum)**.** *Under Assumptions 1, 2 and 4, the map $\vartheta \mapsto \lambda_1(\vartheta)$ is locally Hölder continuous on the set where W_h is nonsingular; and if λ_1 is simple, it is C^1 there, with $\partial\lambda_1/\partial\vartheta$ given by an integral against the stationary measure on $\mathcal{X} \times \mathbb{R}^{d_u} \times \mathbb{P}^{d_h-1}$.*

Two limitations of Figure 10 should be kept in view before it is cited as evidence for anything stronger. First, smoothness here is a visual judgment: we tried to automate it and could not — a discontinuity score built from first or second differences fires on merely-curved-but-continuous stretches whenever the curvature is non-uniform along the path, which it is here, so no calibrated jump test is reported. Second, a discontinuity of Lyapunov exponents in the cocycle would not generically be visible on

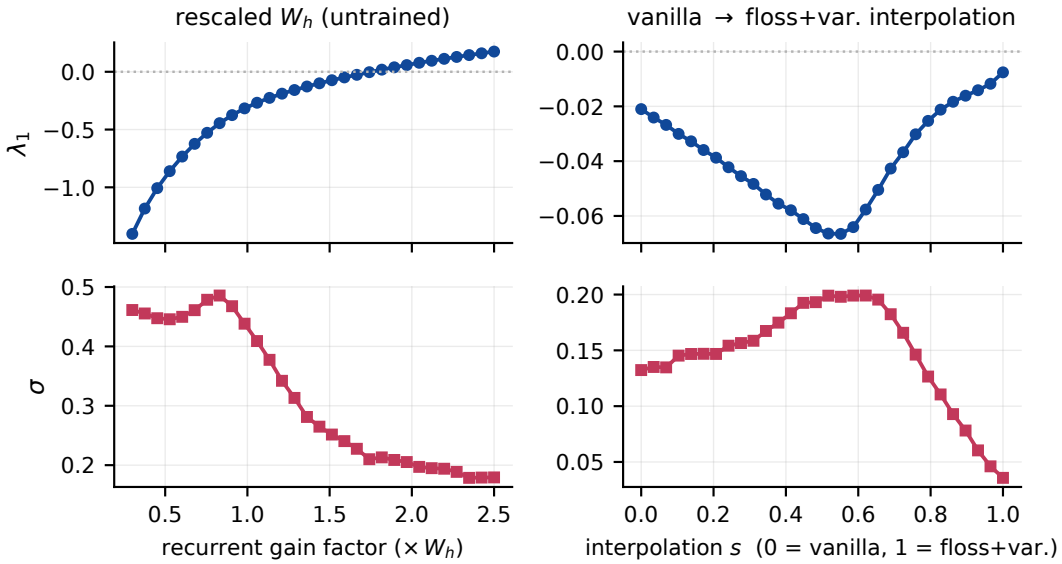

Figure 10: **Numerical probe of Conjecture E.1.** λ_1 (top) and σ (bottom) along two one-parameter paths through weight space. *Left:* a single fixed W_h rescaled continuously, so no fresh randomness enters at any point on the path. *Right:* linear interpolation between the trained `vanilla` and `floss+var` (k=5, $\beta^* = 30$) checkpoints. All four curves are smooth at this resolution, with no visible jump; this is evidence for, though of course not a proof of, continuity of $\vartheta \mapsto \lambda_1(\vartheta)$ in the regime where our assumptions hold. The interpolation path is the more interesting of the two: λ_1 and σ are both non-monotone along it, so the straight line between two trained solutions passes through networks that are simultaneously *more* contracting and *more* diffusive than either endpoint.

a one-dimensional slice sampled at 25–30 points; the figure is consistent with the conjecture rather than a test with power against it.

E.6 The β sweep and the horizon ladder

The two grids of this subsection share a purpose: both vary one thing continuously where Section 5.2 varied a preset name. The β sweep varies the strength of the variance penalty at fixed (α, k); the adding ladder varies the horizon the task demands at fixed method. Each is reported with the run-level numbers in a sidecar JSON written by the script that draws the figure.

Figure 3 shows the β sweep on the m^* axis, which is the one a practitioner can act on. Figure 11 shows the same 30 runs on the raw axes, and two details belong with it. Its left panel prints Pearson and Spearman correlations of accuracy against σ ($r = -0.28$, $\rho = -0.28$); both are misleading if read as "accuracy barely tracks σ", because the relationship is an arch and rank or linear correlation are the wrong summaries for a curve that goes up and then comes down. Its right panel is the induced trajectory in the (λ_1, σ) plane, extended two decades past the main grid's own $\beta = 1$ condition. Nothing is excluded, and every one of the 30 runs has $\lambda_1 < 0$. σ falls monotonically across the entire sweep, including well past the accuracy peak; it is only once both coordinates are combined into m^* that the far end reverses (Figure 3), which is why the design quantity and not σ is what Section 5.1 reports. All numbers, including the per-step monotonicity in units of seed standard deviation and the plateau-membership test, are in the sidecars `figH_beta_sweep_stats.json` and `figH_arch_stats.json`.

E.7 The penalty holds σ down throughout training

F Additional experimental details

Architectures. The psMNIST and adding experiments use two recurrent cells, `vanilla` and `orthogonal`; the nsCIFAR experiments below add GRU (Cho et al., 2014), LSTM (Hochreiter & Schmidhuber, 1997), antisymmetric (Chang et al., 2019), and a selective SSM in the style of Mamba (Gu & Dao, 2023; Gu et al., 2022). Appendix E.3 measures the same five untrained,

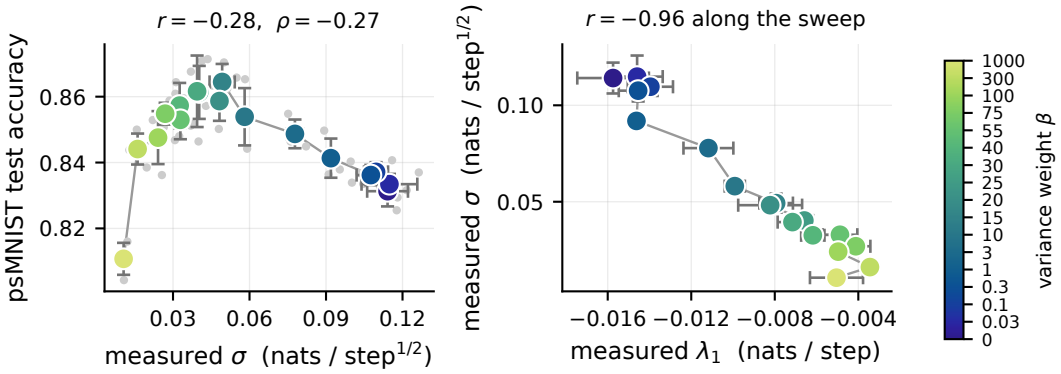

Figure 11: **The β sweep on the raw axes.** The same 30 runs as Figure 3. *Left:* accuracy against measured σ. *Right:* the induced trajectory in the (λ_1, σ) plane, along which σ falls monotonically — including well past the accuracy peak, so the two quantities are not reparametrizations of one another even along this single one-dimensional path.

plus a time-invariant diagonal SSM. `vanilla` is equation 3 with $\phi = \tanh$, $d_h = 128$, and W_h initialized $\mathcal{N}(0, g^2/d_h)$ at gain $g = 1$ unless stated otherwise. `orthogonal` parameterizes $W_h = \rho \exp(A - A^\top)$ with A free and $\rho = 1$, so the recurrence is exactly orthogonal at every step of training; this is the pointwise-constrained baseline of Table 1. All cells read out through a single linear layer from the final hidden state (from every step after the delimiter, for the copy task).

Tasks. *psMNIST*: the 784 pixels of each MNIST image presented one at a time ($d_u = 1$) in a fixed random permutation, 10-way classification by cross-entropy, 5000 images held out from the training split for validation, test accuracy reported on the standard test split. *Adding* (Hochreiter & Schmidhuber, 1997): two input channels of length T, a value channel uniform on $[0, 1]$ and an indicator channel marking two positions, target the sum of the two marked values, mean-squared error; the always-predict-1.0 baseline has MSE ≈ 0.167, and we call a run successful at test MSE ≤ 0.05.

Noise-padded CIFAR-10. Each CIFAR-10 image (Krizhevsky, 2009) is presented row-wise, 32 steps of 96 input dimensions (three channels, per-channel normalized), followed by 968 steps of i.i.d. standard Gaussian noise drawn independently per example rather than shared across a batch, for a fixed sequence length of 1000; solving the task means carrying gradient across the whole padded length. The task and this padding length follow Chang et al. (2019) and Rusch et al. (2022). Seven cells are trained at $d_h = 128$: `vanilla`, `orthogonal`, GRU (87,818 parameters), LSTM at default initialization and at the chrono initialization of Tallec & Ollivier (2018) with $t_{\max} = 1000$ (both 116,490 parameters), `antisymmetric`, and the selective SSM of Appendix E.3 (26,250 parameters); the remaining cells have 30,090 parameters each. Three seeds per condition, 75 runs.

Table 5 reports test accuracy and the trained spectrum at the condition that learns best for each cell: flossing for the GRU and the default-initialized LSTM, no penalty for the rest, since flossing alone already centers their drift and the variance term adds nothing further. `vanilla` and `orthogonal` fail identically and at chance, with drift an order of magnitude away from zero under every penalty setting we tried — the drift failure discussed in Section 5. GRU and chrono-initialized LSTM learn once their drift is centered. The default-initialized LSTM also centers its drift under flossing, but too slowly: its usable horizon is still below 968 at epoch 30, and training ends before it catches up, which we use as the necessary-but-not-sufficient case in Section 5. `antisymmetric` and the selective SSM sit near zero drift from initialization, with no penalty applied, and their accuracy plateaus below the gated cells; for the SSM this reads as an expressivity limit rather than a memory one, since a near-linear map of the input leaves no room for the interactions the task rewards, though we did not test this reading directly. We omit σ for the selective model: at the burn-in used for the trained-checkpoint measurement elsewhere in this appendix its estimate is inflated roughly fourfold for diagonal cells (Appendix E.3), and correcting it is a longer diagnostic run rather than a new training run.

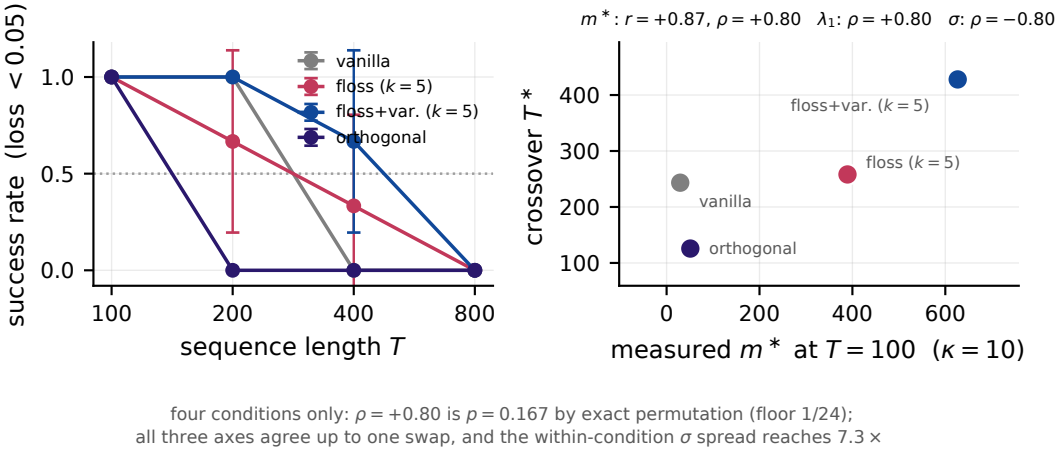

Figure 12: **The horizon ladder brackets each method's crossover, but does not discriminate between candidate axes.** The adding problem at $T \in \{100, 200, 400, 800\}$, four methods, three seeds each (48 runs). *Left:* success rate, the fraction of seeds reaching test loss below 0.05, against T; every method solves $T = 100$ and none solves $T = 800$, so the ladder brackets each crossover, and the variance-aware objective has the latest of the four. *Right:* the crossover T^*, where mean test loss crosses that same threshold, against a measured predictor read at $T = 100$ — the only rung all four methods solve. Three limitations, all of which point the same way. *(i)* With four conditions an exact permutation test over their 24 orderings returns $p = 0.042$, and no arrangement of four could return less than $1/24$: a perfect ordering is the only result this design could have called significant. *(ii)* The predictor is read on networks at a horizon far below the one whose failure it is meant to explain, and it swings by a factor of fifteen across the ladder for the plain recurrence. *(iii)* Most seriously, the per-seed σ within the `vanilla` condition alone spans $\{0.032, 0.234, 0.166\}$, a factor of 7.3 — larger than the between-condition differences any ranking here rests on. We therefore report the bracketing and decline to read a winning axis off four points. Regenerated under equation 11 and the seed-mean convention, all three candidate axes give the same rank ordering up to one adjacent swap ($\rho = +0.80$ for m^* and λ_1, -0.80 for σ), and m^* is ahead only on the linear correlation ($r = +0.87$ against $+0.74$ and -0.39). An earlier version of this caption reported $\rho = +1.00$ for m^* against -0.80 for σ and concluded that m^* was the predictive axis. That was an artifact twice over: the then-current $m^* = (\log \kappa/\sigma)^2$ was a monotone function of σ alone, so the two rank correlations are forced to be exact negatives, and the apparent difference came from averaging m^* over seeds while averaging σ over seeds separately — Jensen, driven by the 7.3× spread in (iii). The left panel appears in the main text as Figure 3 (right); the right panel is kept here because, as (i)–(iii) say, it cannot separate the candidate axes and would read in the main text as a result rather than as the null it is.

Table 5: **Seven cells on noise-padded CIFAR-10** ($N = 1000$), test accuracy and the trained spectrum, three seeds. Reported at the condition that learns best per cell (see text); †chance is 0.10.

	parameters	test acc.	λ_1	σ
vanilla	30,090	$0.096^\dagger$	-0.448	0.145
orthogonal	30,090	$0.100^\dagger$	-0.457	0.109
GRU, flossed	87,818	0.563	0.000	0.012
LSTM, chrono	116,490	0.521	0.000	0.010
LSTM, default, flossed	116,490	$0.107^\dagger$	-0.003	0.021
antisymmetric	30,090	0.371	0.002	0.012
selective SSM	26,250	0.359	-0.001	—

At the same sequence length, Rusch et al. (2022) report 11.6% for a default-initialized LSTM, 43.8% for a GRU, 55.9% for a chrono-initialized LSTM, and 60.5% for their own LEM cell, training for 120 epochs. Our default-initialized LSTM is close to theirs (10.7% against 11.6%), our GRU reaches 56.3% once flossed against their unflossed 43.8%, and our chrono-LSTM reaches 52.1% in 60 epochs against their 55.9% in 120.

Sub-horizon learning and dynamic range tolerance. While $m^* \geq H_{\text{task}}$ acts as a graded necessary condition, two configurations in the benchmarks learn despite $m^* < H_{\text{task}}$. First, on the adding problem at $T = 100$, vanilla RNNs solve the task ($8/8$ seeds) with $m^* \approx 14$–30. Second, on nsCIFAR ($H_{\text{task}} = 968$), the antisymmetric network reaches 0.371 test accuracy with $m^* \approx 630$–750. Both cases reflect the nominal choice of $\kappa = 10$ in equation 11. In practice, training with Adam

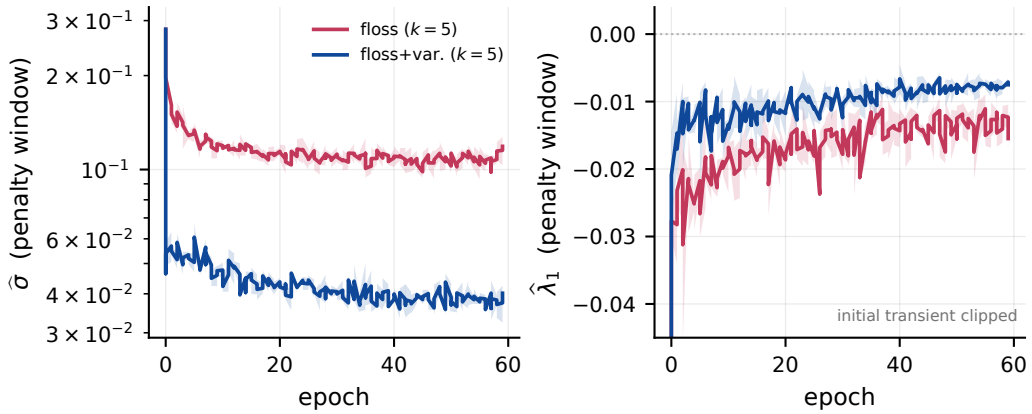

Figure 13: $\widehat{\sigma}$ **and** $\widehat{\lambda}_1$ **through training**, mean over three seeds with a ± 1 s.d. band, for the two settings that differ only in the variance term ($\alpha = 1$, $k = 5$, $\beta = 0$ against $\beta^* = 30$). "The penalty controls σ" would be a weak claim if the two runs merely ended in different places; they do not. $\widehat{\sigma}$ separates within the first epoch (0.054 against 0.197) and stays a factor of about three apart for the whole budget, and the variance-aware run also holds $\widehat{\lambda}_1$ closer to zero throughout. These are the training-window estimates — the $m = 128$ penalty window on 32 trajectories — so they carry the finite-horizon bias of Section **??** and are not directly comparable to the long double-precision measurements of Table 1; what they show is the *shape* of the trajectory. The unregularized and orthogonal baselines cannot appear here: with $\alpha = \beta = 0$ no penalty is computed, so no per-step $\widehat{\sigma}$ is logged. The initial transient ($\widehat{\lambda}_1 \approx -0.33$ at initialization) is clipped in the right panel.

and gradient norm clipping (1.0) tolerates coordinate-wise gradient dynamic range compressions of 10^2–10^3 ($\log \kappa \approx 4.6$–6.9). Re-evaluating equation 11 at $\kappa = 10^3$ scales m^* up by a factor of 3–9, shifting the usable horizon of the antisymmetric cell to $m^* \approx 1800 > 968$ and the adding RNN to $m^* \approx 110 > 100$. Hence, m^* represents a comparative scale rather than a sharp cutoff.

Optimization. Adam, learning rate 10^{-3} with a cosine schedule, batch size 512, gradient-norm clipping at 1.0, 60 epochs, no weight decay. Five seeds (0–4) for Table 1 and the β sweep, three seeds (0–2) elsewhere; every reported mean is over completed seeds and every error bar is stated in the caption as either s.d. or s.e.m.

The penalty and its estimator. The differentiable QR recursion of Vogt et al. (2022) is run on a window of $m = 128$ steps with a 64-step frame burn-in, on 32 of the 512 trajectories in each batch, with the full (non-truncated) frame gradient. The finite-time exponents are the means of the accumulated $\log R_t[i,i]$, and $\widehat{\sigma}^2$ is equation 43 applied to the same f_t at block length $\ell = \lfloor \sqrt{m} \rfloor$ with non-overlapping blocks. The training objective is $\mathcal{L}_{\text{task}} + \alpha \sum_{i \le k} \lambda_i^2 + \beta \widehat{\sigma}^2$.

The frame gradient is not an approximation. Implementations of a Lyapunov penalty differ in whether they backpropagate through the QR frame or detach it, and the detached form is the cheaper and more common one. It is not an approximation of the other. Measured in our setting ($d_h = 128$, $k = 5$, a 128-step window with a 64-step burn-in on 32 trajectories, six random initializations), the cosine between the two parameter gradients is 0.694 ± 0.025 and the detached one is 0.707 ± 0.021 times the length of the full one: the same sign, but a different direction and a systematically smaller step. All runs in this paper use the full frame gradient. We report this because it is a fact about gradient flossing as commonly implemented rather than about our own method, and because a cosine of 0.7 is close enough to 1 to look like agreement in a plot and far enough to change what is being optimized.

Measurement after training. λ_1 and σ are re-measured on the trained checkpoint in float64 on 16 trajectories of $20{,}000$ steps with a 200-step frame burn-in, independently of the trajectories used in training. For the diagonal cells (selective and LTI SSM) the frame burn-in is 10^4 steps instead, since 200 steps is not enough for the frame to align with the top channel (Appendix E.3). Unless stated otherwise the measurement uses the task input law; Table 3 reports both laws. The usable horizon m^* is equation 11 at $\kappa = 10$ and $\alpha = 0.05$, evaluated at the seed-mean (λ_1, σ).

Cost of the variance term. Measured by `lyapunov_rnn.bench` on one CPU with $d_h = 128$, $T = 784$, $k = 5$, a 128-step penalty window on 32 trajectories, and 12 timed steps after two warm-up steps. At the batch size used for training ($B = 512$): 0.277 s/step unregularized, 0.390 with flossing ($+41\%$), 0.388 with the variance term added (-0.5% relative to flossing). At $B = 128$: 0.098, 0.198 ($+103\%$), 0.205 ($+3.4\%$). The variance term is therefore free to within timing noise, and what a variance-aware objective actually costs is the QR recursion it shares with gradient flossing. On GPU the reduction is if anything cheaper relative to the recursion, so these CPU figures are a conservative statement of the overhead.

Compute and reproducibility. The training results rest on 478 completed runs, all 60 epochs at batch 512: 185 on psMNIST, 194 on the adding problem and 99 on noise-padded CIFAR-10, with three to eight seeds per condition. Each run used a single NVIDIA A100 or V100 GPU (San Diego Supercomputer Center's TSCC cluster), totaling approximately 214 GPU-hours (168 on A100, 46 on V100). The diagnostics of Appendix E require no training and each costs minutes on one CPU. Every number in Section 5 is emitted from `runs/*/results.csv` by the same plotting modules that draw the figures, with per-figure sidecar JSONs, so prose and panels cannot drift apart.

How Figure 1 was produced. *Panel (a).* The network is the untrained `vanilla` reference ($d_h = 128$, Gaussian initialization, the same checkpoint used for E2 and E4) with its recurrent matrix multiplied by a single scalar, found by bisection (12 iterations, each evaluating λ_1 on 8 trajectories of 6000 steps) until the measured top exponent vanished. The factor is 1.7590 and the residual is $\lambda_1 = +0.0007$; because only that one scalar moves, the fan that remains cannot be an artifact of redrawing W_h. *Panel (b)* uses the trained `floss_topk` and `floss_var` ($\beta^* = 30$) checkpoints at seed 0, which differ in the variance term and in nothing else. In every panel the pair (λ_1, σ) that draws the envelope is measured afterwards on a separate run of 16 trajectories $\times$ 8000 steps with its own seed, σ by the batched-means estimator equation 43 at block length $\lfloor\sqrt{m}\rfloor = 89$; the paths and the marginal histograms are then 512 further trajectories drawn with a third seed, of which 160 are drawn as curves and all 512 enter the histogram. The input law is i.i.d., under which Assumption 1 holds verbatim; Figure 2 is the corresponding statement for the psMNIST stream, where it does not. Each panel is one call to `diagnostics fan --run <ckpt> --law iid [--calibrate]` followed by `plotting/fig_fan.py`, and takes seconds on CPU.