Dimension Bounds for Contractive Reservoir Computing from Input Entropy

Pradeep Singh*
Department of Computer Science
and Engineering
Indian Institute of Information
Technology Surat
Surat, India
pradeep.singh@iiitsurat.ac.in

Kishore Babu Nampalle
Department of Computer Science
and Engineering
Indian Institute of Information
Technology Ranchi
Ranchi, India
nkishore@iiitranchi.ac.in

Balasubramanian Raman
Department of Computer Science
and Engineering
Indian Institute of Technology
Roorkee
Roorkee, India
bala@cs.iitr.ac.in

Abstract

Echo State Networks (ESNs) are designed to be *stable*—the echo state property makes the reservoir state a well-defined function of the input history—but it remains unclear how *geometrically rich* the reachable reservoir states are and what controls that richness. We give a partition-free, fractal-geometric theory of contractive reservoirs by viewing a quantized-input ESN as a contractive random dynamical system and, equivalently, an iterated function system (IFS). This identifies the reachable set as the IFS attractor and the stationary reservoir-state law under i.i.d. inputs as the unique invariant IFS measure. We prove entropy–contraction (entropy–Lyapunov) bounds on the intrinsic dimension of this measure, and in a conformal/similarity regime with standard separation conditions we obtain sharp closed-form dimension laws of the form "input entropy divided by contraction." These results yield a quantitative design principle: input entropy and contraction impose a necessary threshold for full-dimensional state representations, providing guidance for tuning leak, recurrent gain, and input scaling.

1 Introduction

Echo State Networks (ESNs) are typically justified by *stability*: under a contractivity condition, the reservoir satisfies the echo state property (ESP), meaning that the current state is a well-defined function of the input history and remote inputs are exponentially forgotten [26, 32, 31]. Yet stability alone does not explain a second, equally practical phenomenon: contractive reservoirs can generate surprisingly *rich* feature representations for linear readouts. This paper addresses a basic question that lies between dynamical-systems theory and representation learning: *how geometrically rich is the set (and distribution) of reservoir states that can actually be reached under admissible inputs, and what controls that richness?*

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Recent years have strengthened both the theory and practice of reservoir computing beyond classical ESP-style stability analyses. On the theory side, ESNs and related reservoir systems have been shown to be *universal* approximators of fading-memory input–output maps under increasingly general assumptions [21, 18, 19], and these results have been complemented by differentiable formulations that enable gradient-based tuning of reservoir mechanisms [22] and by statistical learning-theoretic guarantees controlling finite-sample generalization of readouts trained on reservoir features [17]. In parallel, architectural developments such as deep/stacked reservoirs clarify how structural choices create multi-timescale internal representations [15], while "next-generation" reservoir computing reinterprets reservoirs as explicit nonlinear feature maps with strong empirical performance in dynamical forecasting [16]; related analyses connect reservoir representations to systematic expansions such as discrete-time signatures [7]. Meanwhile, the broader sequence-modeling literature has also emphasized stability and Lipschitz constraints as a route to trainable and robust recurrent architectures [33, 12]. Despite this progress, the *intrinsic geometry* of stable reservoir state representations—the dimension of the reachable set and the stationary state distribution induced by stochastic inputs—remains largely implicit.

Why geometry is the missing invariant. From a representation-learning viewpoint, a contractive reservoir is a map from input histories to points in $\mathbb{R}^N$; what the readout "sees" is not the full ambient space but the subset of states actually realized under admissible inputs. Two reservoirs can share the same ESP-style fading-memory guarantee [26, 32] and yet induce radically different state-space geometry: one may explore a full-dimensional region, while another collapses onto a thin (possibly fractal) subset whose intrinsic degrees of freedom are far smaller than N. This distinction is invisible to approximation-theoretic universality statements and complementary to capacity-style summaries [8], and quantifies the growth of quantized-state entropy as resolution increases [37, 40]. This geometric quantity complements covariance rank and task-specific readout capacity; it does not by itself determine either. Our goal is therefore to make the geometry explicit by deriving intrinsic dimension laws from first principles, without choosing partitions or training any readout.

We propose a partition-free approach based on *fractal state geometry*. When inputs are finite or quantized, a contractive ESN induces a finite family of contracting maps $\{F_s\}_{s\in\Sigma}$ on the state space. This places contractive reservoirs within the classical theory of iterated function systems (IFS) [25, 2] and, equivalently, within contractive random dynamical systems (RDS) [1]. Two intrinsic objects emerge: (i) the *reachable state set* K, the collection of all states that can appear at a fixed time as the input past varies, and (ii) the *stationary state distribution* μ induced by a stationary input process. In the contractive regime, K is naturally identified with the IFS attractor, and for i.i.d. inputs μ is the unique invariant IFS measure solving a pushforward fixed-point equation [25, 14]. These objects enable an intrinsic quantitative notion of reservoir "richness" via fractal dimensions of K and μ.

Our main theoretical message is an entropy–contraction principle: *the geometric complexity of reservoir states is controlled by the input entropy rate divided by a contraction exponent*. Concretely, letting $h(U)$ be the entropy rate of the input process [39, 6] and letting $\chi < 0$ denote either the uniform exponent or a finite deterministic pathwise exponent satisfying Theorem 3.1, we prove an information-dimension upper bound of the form $\overline{\dim}_I(\mu) \le \min\left\{N, \frac{h(U)}{|\chi|}\right\}$, which parallels classical entropy–dimension relations in smooth and random dynamical systems [41, 29]. Importantly, our bound is derived directly from the ESP geometry: conditioning on a length-n input block restricts the current state to a set of diameter $\approx e^{n\chi}$, while the number of distinguishable blocks grows like $e^{nh(U)}$. As a result, contraction collapses the representation and input entropy expands it. At a fixed input entropy rate, a desired dimension requires sufficiently weak contraction; at a fixed contraction exponent, low entropy limits the upper information dimension. These are necessary constraints, with attainability depending on the reservoir geometry.

Contributions (informal). (i) We formalize contractive ESNs as finite IFS/RDS objects and identify the reachable set K with the IFS attractor, and (for i.i.d. inputs) the stationary state distribution μ with the unique invariant IFS measure [25, 1]. (ii) We introduce *reservoir dimension* as a partition-free measure of representational richness (Hausdorff/Minkowski dimension for K, and information dimension for μ) [14, 37, 40]. (iii) We prove entropy–contraction (entropy–Lyapunov) dimension bounds that connect $h(U)$ and contraction exponents to $\dim_I(\mu)$; all formal proofs are deferred to the supplementary material.

2 Setup: Contractive Reservoirs as IFS / Random Dynamical Systems

2.1 Contractive ESN model and echo-state canonical solution

Let Σ be a finite input alphabet (e.g., a quantization of a continuous input), and let $(U_t)_{t\in\mathbb{Z}}$ be a stationary stochastic process on Σ. We consider a leaky ESN update on a compact state space $X \subset \mathbb{R}^N$:

$$x_{t+1} = F_{U_t}(x_t) \text{ with } F_s(x) := (1-\alpha)x + \alpha\,\phi(Wx + W_{\text{in}}s + b), \tag{1}$$

where $\alpha \in (0,1]$ is the leak, $W \in \mathbb{R}^{N\times N}$, $W_{\text{in}} \in \mathbb{R}^{N\times d}$, $b \in \mathbb{R}^N$, and ϕ is applied coordinatewise. We work in the standard *contractive (ESP) regime* [26, 32]:

Assumption 2.1 (Forward invariance and uniform contractivity)**.** *There exists a nonempty compact set $X \subset \mathbb{R}^N$ such that $F_s(X) \subseteq X$ for all $s \in \Sigma$, and there exists $L \in (0,1)$ such that $\|F_s(x) - F_s(y)\| \le L\|x-y\|\ \forall x,y \in X,\ \forall s \in \Sigma$. For* (1)*, a sufficient condition is $(1-\alpha)+\alpha\|W\| < 1$ when ϕ is 1-Lipschitz [26, 32].*

Under Assumption 2.1, every bi-infinite input trajectory admits a unique bi-infinite state trajectory (the echo state property). A convenient formulation is the *pullback/canonical* state construction from the infinite past, standard in contractive skew products and random dynamical systems [1]:

Theorem 2.2 (Echo state canonical solution (contractive case))**.** *Assume Assumption 2.1. For any realization of the input process $(U_t)_{t\in\mathbb{Z}}$ and any $t \in \mathbb{Z}$, the limit*

$$X_t := \lim_{n\to\infty} F_{U_{t-1}} \circ F_{U_{t-2}} \circ \cdots \circ F_{U_{t-n}}(x^\star) \tag{2}$$

exists for any $x^\star \in X$, is independent of $x^\star$, and defines the unique bounded solution to (1)*. Moreover, the map from past inputs to the present state is exponentially stable: if two input histories agree on the last m symbols, then $\|X_t - \widetilde{X}_t\| \le L^m \operatorname{diam}(X)$.*

Theorem 2.2 implies that, in the contractive regime, the reservoir state is a measurable *filter* of the input past with a quantitative fading-memory modulus [26, 32, 21]. This is the stability backbone on which we build a geometric complexity theory.

2.2 IFS attractor and stationary measure

Because the input alphabet is finite, (1) induces a finite family of contractions $\{F_s\}_{s\in\Sigma}$ on X. This family defines an iterated function system (IFS) [25, 2]. Let $\mathcal{K}(X)$ denote the nonempty compact subsets of X with the Hausdorff metric d_H. Define the Hutchinson operator $\mathcal{H}(A) := \bigcup_{s\in\Sigma} F_s(A),\ A \in \mathcal{K}(X)$. Since each F_s is L-Lipschitz, $\mathcal{H}$ is a contraction on $(\mathcal{K}(X), d_H)$ with Lipschitz constant L [25]. Hence it has a unique fixed point:

Theorem 2.3 (IFS attractor)**.** *Under Assumption 2.1, there exists a unique nonempty compact set $K \subseteq X$ such that $K = \mathcal{H}(K) = \bigcup_{s\in\Sigma} F_s(K)$. Moreover, for any nonempty compact $A \subseteq X$, $\mathcal{H}^n(A) \to K$ in Hausdorff metric as $n \to \infty$.*

We interpret K as the *reachable reservoir state set*: it is the collection of all present states attainable by varying the infinite past input. Formally, let $\Sigma^{\mathbb{Z}_{<0}}$ denote one-sided infinite pasts. The pullback construction (2) induces a map $\Psi : \Sigma^{\mathbb{Z}_{<0}} \to X$ sending an input past to X_0. Then $K = \Psi(\Sigma^{\mathbb{Z}_{<0}})$: the coding map is continuous by the fading-memory estimate, and its domain is compact. A proof is given in Appendix A.3.

Next, when the input process is stochastic, we obtain a canonical stationary state distribution. For i.i.d. inputs with law p on Σ, define the Markov operator on probability measures $\mathcal{P}(X)$: $\mathcal{T}_p(\nu) := \sum_{s\in\Sigma} p(s)\,(F_s)_{\#}\nu$, where $(F_s)_{\#}\nu$ is the pushforward measure. Under contractivity, $\mathcal{T}_p$ is a contraction in Wasserstein-1 distance, yielding a unique invariant measure [25, 11]:

Theorem 2.4 (Stationary IFS measure (i.i.d. inputs))**.** *Assume Assumption 2.1 and i.i.d. inputs $U_t \sim p$. Then there exists a unique probability measure $\mu_p \in \mathcal{P}(X)$ satisfying*

$$\mu_p = \mathcal{T}_p(\mu_p) = \sum_{s\in\Sigma} p(s)\,(F_s)_{\#}\mu_p. \tag{3}$$

If X_0 is distributed according to μ_p and is independent of $(U_t)_{t\ge 0}$, and $X_{t+1} = F_{U_t}(X_t)$, then (X_t) is stationary with marginal law μ_p, and $\operatorname{supp}(\mu_p) \subseteq K$.

Remark (general stationary inputs). For stationary ergodic (not necessarily i.i.d.) inputs, Theorem 2.2 still yields a unique stationary state process (X_t) as a measurable function of the input past [1]. In Section 3 we formulate entropy-rate dimension bounds in terms of the entropy rate $h(U)$ of the general input process. When i.i.d. the stationary distribution also satisfies the explicit fixed-point relation (3); for Markov inputs one can formulate an analogous invariant measure for the skew product on $\Sigma \times X$ [1].

2.3 Notions of dimension used in this paper

We use two complementary notions of geometric complexity: a dimension of the reachable set K, and a dimension of the stationary state distribution μ.

Set dimension. For a bounded set $A \subset \mathbb{R}^N$, let $N_\varepsilon(A)$ be the minimal number of Euclidean balls of radius ε needed to cover A. The (upper) Minkowski dimension is $\overline{\dim}_M(A) := \limsup_{\varepsilon \downarrow 0} \frac{\log N_\varepsilon(A)}{\log(1/\varepsilon)}$. We will also refer to Hausdorff dimension $\dim_H(A)$ [14]. In general $\dim_H(A) \le \overline{\dim}_M(A)$.

Measure dimension. Let μ be a Borel probability measure on $\mathbb{R}^N$, and let $Q_\varepsilon : \mathbb{R}^N \to \mathbb{Z}^N$ denote uniform quantization at resolution ε (e.g., mapping a point to the index of the ε-cube in a fixed axis-aligned grid). The (upper) *Rényi information dimension* of μ is defined by [37, 40] $\overline{\dim}_I(\mu) := \limsup_{\varepsilon \downarrow 0} \frac{H\left(Q_\varepsilon(Z)\right)}{\log(1/\varepsilon)}$, $Z \sim \mu$, where $H(\cdot)$ denotes Shannon entropy (natural logs). When the limsup is a limit, the common value is the information dimension $\dim_I(\mu)$. Information dimension is particularly natural here because it measures how the discrete entropy of a quantized state grows as resolution increases, making it directly aligned with the "effective degrees of freedom" of the reservoir representation.

Exact dimensionality. A measure μ is called *exact dimensional* if the pointwise (local) dimension exists and is constant μ-a.e.; in that case the constant equals $\dim_H(\mu)$ and coincides with $\dim_I(\mu)$ under mild regularity [14, 29]. In later sections (and in the supplementary material) we identify regimes where the stationary measure μ is exact dimensional and dimension bounds become equalities.

A multi-resolution picture: input cylinders induce state coverings. The IFS/RDS viewpoint is useful because it turns "fading memory" into a geometric, scale-by-scale statement. Fix a time (say $t = 0$). For each length-n input past block $u_{-n:-1} \in \Sigma^n$, consider the set of all possible present states consistent with that block when the earlier past is arbitrary; by Assumption 2.1, this conditional set has diameter at most $L^n \mathrm{diam}(X)$, because any two trajectories that share the last n symbols must contract by at least L^n before reaching time 0. As n grows, these cylinder-conditioned sets form a nested family of coverings of the reachable set K and, under stochastic inputs, of the typical support of the stationary measure μ. Section 3 formalizes the resulting counting argument: the number of typical input blocks grows like $\exp(nh(U))$ while the corresponding state sets shrink like $\exp(n\chi)$, yielding an entropy–contraction control of intrinsic dimension [41, 29].

3 Entropy–Lyapunov Dimension Bounds

This section connects the stochastic complexity of inputs to the geometric complexity of reservoir states. The overarching principle is classical: entropy creates variety while contraction collapses it, so their ratio controls dimension [41, 29]. In our setting, the key technical mechanism is simple: conditioning on a finite input block restricts the current state to a small set (by contraction), and the number of typical input blocks grows exponentially at the entropy rate.

Entropy rate of the input process. Let $U = (U_t)_{t \in \mathbb{Z}}$ be a stationary process on the finite alphabet Σ. For $n \ge 1$, write $U_{0:n-1} := (U_0, \dots, U_{n-1})$. The block entropy is $H(U_{0:n-1})$, and the entropy rate is [39, 6] $h(U) := \lim_{n \to \infty} \frac{1}{n} H(U_{0:n-1})$, which exists for stationary finite-alphabet processes. For i.i.d. inputs with symbol law p, $h(U) = H(p) := -\sum_{s \in \Sigma} p(s) \log p(s)$. For a stationary Markov chain with transition matrix P and stationary distribution π, one has $h(U) = -\sum_{i \in \Sigma} \pi_i \sum_{j \in \Sigma} P_{ij} \log P_{ij}$ [6]. Throughout we use natural logarithms, so entropies are in nats.

Lyapunov / contraction exponents for the reservoir. Define the n-step composition along a realized input block: $\Phi(n; U_{0:n-1}) := F_{U_{n-1}} \circ \cdots \circ F_{U_1} \circ F_{U_0}$. Let $\mathrm{Lip}(G)$ denote the global

Lipschitz constant of a map G. Under Assumption 2.1, $\mathrm{Lip}\big(\Phi(n;U_{0:n-1})\big) \le L^n$. This yields a uniform contraction exponent $\chi_{\mathrm{unif}} := \log L < 0$.

It is often useful to track *pathwise* contraction, which can be strictly stronger than L^n. Define the random variables $\Lambda_n := \log \mathrm{Lip}\big(\Phi(n;U_{0:n-1})\big)$. The sequence is subadditive: $\Lambda_{n+m} \le \Lambda_n \circ \sigma^m + \Lambda_m$. For stationary ergodic inputs, if $\Lambda_n \in L^1$ for every n and $\inf_{n\ge 1} n^{-1}\mathbb{E}[\Lambda_n] > -\infty$, Kingman's theorem gives a finite deterministic limit

$$\chi := \lim_{n\to\infty} \frac{1}{n}\Lambda_n \tag{4}$$

almost surely and in L^1 [28]. Uniform contractivity gives $\chi \le \log L < 0$. The refined bound below only assumes a finite deterministic almost-sure limit; L^1 convergence is not needed.

Differentiable case (Lyapunov exponents). If each F_s is differentiable and $\log^+ \|DF_{U_0}(X_0)\|$ is integrable under stationarity, one may define a (top) Lyapunov exponent via Oseledets' theorem [36], and relate it to χ; we keep the Lipschitz-based formulation (4) since it applies directly to standard ESNs under global Lipschitz control.

Main bound: dimension is controlled by entropy divided by contraction. Let X_t be the canonical state process from Theorem 2.2 driven by the stationary input process U, and let μ denote the stationary marginal distribution of X_0.

Theorem 3.1 (Entropy–contraction (entropy–Lyapunov) upper bound)**.** *Assume Assumption 2.1 and let U be stationary with entropy rate $h(U)$. Let μ be the stationary law of the canonical state X_0. (i) (Uniform form) The upper information dimension satisfies $\overline{\dim}_I(\mu) \le \min\Big\{N,\ \frac{h(U)}{|\log L|}\Big\}$. (ii) (Refined pathwise form) If additionally Λ_n/n converges almost surely to a deterministic constant $\chi \in (-\infty, 0)$, then $\overline{\dim}_I(\mu) \le \min\Big\{N,\ \frac{h(U)}{|\chi|}\Big\}$.*

Corollary (i.i.d. inputs). If U_t are i.i.d. with law p on Σ, then $h(U) = H(p)$ and $\overline{\dim}_I(\mu) \le \min\Big\{N,\ \frac{H(p)}{|\log L|}\Big\}$. At fixed $L < 1$, sufficiently low input entropy therefore bounds the state information dimension independently of reservoir width.

Remark (reachable-set dimension). A related bound applies to the reachable set K. Since K is covered by at most $|\Sigma|^n$ images $\Phi(n;w)(X)$ of diameter $\le L^n \mathrm{diam}(X)$ (one for each word $w \in \Sigma^n$), we obtain $\overline{\dim}_M(K) \le \min\Big\{N,\ \frac{\log|\Sigma|}{|\log L|}\Big\}$, a purely combinatorial–geometric upper bound consistent with classical IFS covering arguments [25, 14]. In regimes with weak overlaps (e.g., separation conditions), such bounds can become sharp and yield closed-form dimension laws; we return to this in subsequent sections and the supplementary material.

4 Exact Dimension Laws in the Similarity / Conformal Regime

Section 3 established that, for any contractive reservoir, the geometric complexity of the stationary state distribution μ is upper bounded by an "entropy divided by contraction" quantity (Theorem 3.1). This section explains when such bounds are *sharp* and admit *closed-form* expressions. The key additional ingredient is a classical one: when the reservoir maps are *similarities* and their images satisfy the separation hypothesis below, the reachable set K and the stationary measure μ fall into the well-studied theory of self-similar fractals and self-similar measures [35, 25, 14]. In reservoir terms, this corresponds to a regime where the reservoir contracts nearly isotropically and where distinct input symbols induce sufficiently distinct "translations" so that different branches of the dynamics do not overlap excessively.

4.1 Similarity reservoirs and separation conditions

We introduce a tractable, mathematically crisp subclass that yields exact formulas and serves as a reference point for general ESNs.

Assumption 4.1 (Similarity (conformal) reservoir)**.** *Each fiber map is a similarity on $\mathbb{R}^N$ with a common contraction ratio $r \in (0,1)$: $F_s(x) = r\,R_s x + b_s,\ s \in \Sigma$, where $R_s \in O(N)$ is orthogonal*

and $b_s \in \mathbb{R}^N$ is a translation vector. We assume there is a compact forward-invariant set X such that $F_s(X) \subseteq X$ for all s (hence Assumption 2.1 holds with $L = r$).

Assumption 4.1 defines a classical self-similar IFS [25, 2, 14]. The reachable set K is its attractor (Theorem 2.3). In general, the Hausdorff dimension of K and the dimension of the stationary measure can drop below their "ideal" values if images overlap. The standard way to rule out pathological overlaps is a separation condition.

Definition 4.2 (Open set condition (OSC) and strong separation (SSC))**.** *Let $\{F_s\}_{s\in\Sigma}$ be an IFS with attractor K. (i) The IFS satisfies the* open set condition (OSC) *if there exists a nonempty bounded open set $O \subset \mathbb{R}^N$ such that $\bigcup_{s\in\Sigma} F_s(O) \subseteq O$ and $F_s(O) \cap F_{s'}(O) = \varnothing$ for $s \neq s'$. (ii) The IFS satisfies the* strong separation condition (SSC) *if $F_s(K) \cap F_{s'}(K) = \varnothing$ for all $s \neq s'$.*

SSC implies OSC, and OSC is sufficient for many exact dimension formulas in the similarity case [25, 14, 38]. In reservoir language, OSC/SSC can be interpreted as guaranteeing that different input symbols lead to distinct "branches" of the state evolution at every scale, preventing uncontrolled aliasing in the reachable set.

Connection to general ESNs. Standard ESN updates with smooth nonlinearities are not exactly similarities; nonetheless, Assumption 4.1 should be viewed as a *conformal reference model* capturing a regime where the reservoir behaves approximately like an isotropic contraction with input-dependent translations. This occurs, for instance, when the Jacobians DF_s have bounded distortion and their singular values are nearly equal ("approximately conformal"). In such regimes one expects entropy–Lyapunov formulas analogous to those stated below, possibly up to bounded-distortion corrections [14, 29]. We focus on the similarity case because it yields explicit closed forms and a clean design principle.

4.2 Closed-form dimension formulas for K and μ

In the similarity regime with separation, the classical "similarity dimension" controls the dimension of the attractor [35, 25].

Definition 4.3 (Similarity dimension)**.** *Under Assumption 4.1, define $s_{\mathrm{sim}} := \frac{\log|\Sigma|}{|\log r|}$. More generally, if F_s have possibly different contraction ratios r_s, then s_{sim} is defined as the unique solution of Moran's equation $\sum_{s\in\Sigma} r_s^{s_{\mathrm{sim}}} = 1$ [35, 14].*

Theorem 4.4 (Exact dimension of the reachable set in the similarity case)**.** *Assume Assumption 4.1 and that the IFS satisfies OSC. Let K be the IFS attractor. Then $\dim_H(K) = \overline{\dim}_M(K) = \min\{N,\ s_{\mathrm{sim}}\}$.*

Next consider stochastic inputs. For i.i.d. inputs $U_t \sim p$ on Σ, the stationary distribution μ from Theorem 2.4 is precisely the self-similar measure associated with weights p [25, 14]. Under OSC, it is exact-dimensional with a closed-form dimension.

Theorem 4.5 (Exact dimension of the stationary measure in the similarity case)**.** *Assume Assumption 4.1 and OSC. Let (U_t) be i.i.d. with symbol distribution p, and let μ be the unique stationary law of X_0. Then μ is exact-dimensional and $\dim_I(\mu) = \dim_H(\mu) = \min\left\{N,\ \frac{H(p)}{|\log r|}\right\}$, where $H(p) = -\sum_{s\in\Sigma} p(s)\log p(s)$.*

Relationship to entropy–Lyapunov bounds. Under Assumption 4.1, the contraction exponent is $\chi = \log r$. Thus Theorem 4.5 shows that the entropy–contraction upper bound of Theorem 3.1 is *tight* within the parameter range compatible with OSC: $\dim_I(\mu) = \min\left\{N,\ \frac{h(U)}{|\chi|}\right\}$ (here $h(U) = H(p)$). Under OSC, $H(p) \le \log|\Sigma| \le N|\log r|$, so the minimum with N is redundant. The next section interprets the bound as a necessary design criterion.

Caveat: overlaps can reduce dimension. Without OSC/SSC, exact overlaps can cause strict dimension drop below the similarity prediction, even in the similarity setting [14]. In reservoir terms, strong overlap corresponds to different input histories collapsing to the same state at many scales. Our general results (Section 3) remain valid without separation and provide robust upper bounds in such cases.

5 A Criticality Principle for Reservoir Design

Reservoir computing practice has long observed that performance often improves as the reservoir approaches the boundary of stability (e.g. spectral radius near 1), a phenomenon commonly framed as operating near a "critical" point [3, 4]. The entropy–contraction bound provides a quantitative necessary condition for full-dimensional representations within the strictly contractive ESP regime. The exact similarity formulas describe when the bound is attained under OSC.

5.1 A necessary threshold for full-dimensional representations

In the similarity regime (Assumption 4.1 with OSC), Theorem 4.5 implies that the intrinsic representation dimension equals

$$d_\star(r,p) \;:=\; \dim_I(\mu) \;=\; \min\Big\{N,\; \frac{H(p)}{|\log r|}\Big\}. \tag{5}$$

More generally, even without OSC, Theorem 3.1 gives the necessary condition

$$\overline{\dim}_I(\mu) = N \implies \frac{H(p)}{|\log r|} \geq N \iff r \geq r_c(p) := \exp\Big(-\frac{H(p)}{N}\Big). \tag{6}$$

This condition is not sufficient: overlaps can reduce the realized dimension. Under OSC, volume packing imposes $|\Sigma| r^N \leq 1$, so $H(p)/|\log r| \leq N$. Theorem 4.5 attains dimension N only when p is uniform and $|\Sigma| r^N = 1$. Thus the OSC formula does not establish a full-dimensional regime beyond this boundary.

Interpretation inside the ESP regime. For $r < r_c(p)$, the bound rules out full-dimensionality; for $r \geq r_c(p)$, the bound alone no longer rules it out. Weakening contraction enlarges the dimension permitted by the bound, without guaranteeing that the realized dimension increases. At a fixed contraction exponent, full-dimensionality is impossible whenever $h(U) < N|\chi|$. For positive input entropy, $r_c(p) < 1$, so this necessary threshold lies within the contractive regime.

Uniform inputs. If p is uniform, then $H(p) = \log|\Sigma|$ and $r_c = |\Sigma|^{-1/N}$. For a fixed alphabet size greater than one, this necessary threshold approaches 1 as N increases. This motivates studying weakly contractive reservoirs, while leaving attainability to the geometry of the particular model.

5.2 Design rules: spectral radius/leak vs input entropy

Outside the exact similarity case, Theorem 3.1 supplies an upper bound $\overline{\dim}_I(\mu) \leq \min\{N, h(U)/|\chi|\}$, using either $\chi = \log L$ or the finite deterministic pathwise exponent specified there. It gives feasibility constraints and motivates parameter choices; it does not guarantee a target dimension. **Rule 1: use the dimension bound as a feasibility criterion.** For $0 < d_{\text{target}} \leq N$, achieving $\overline{\dim}_I(\mu) \geq d_{\text{target}}$ requires $|\chi| \leq h(U)/d_{\text{target}}$. In a similarity proxy this becomes $r \geq \exp(-h(U)/d_{\text{target}})$. The boundary value $r_{\text{target}} := \exp(-h(U)/d_{\text{target}})$ is a candidate tuning target when it lies in $(0,1)$; overlaps and anisotropy may yield a smaller realized dimension. Taking $d_{\text{target}} = N$ recovers (6).

Rule 2: translate r_{target} into ESN parameters via Lipschitz control. For the leaky ESN update $F_s(x) = (1-\alpha)x + \alpha\,\phi(Wx + W_{\text{in}}s + b)$ with ϕ 1-Lipschitz, a standard sufficient bound is $\text{Lip}(F_s) \leq (1-\alpha) + \alpha\|W\| \Rightarrow L := (1-\alpha) + \alpha\|W\| < 1$ to ensure ESP [26, 32]. A conservative design is therefore to set $L \approx r_{\text{target}}$ while keeping $L < 1$ with margin. If W is normal (e.g. orthogonally diagonalizable) or explicitly chosen orthogonal/unitary up to a gain factor, then $\|W\|_2 = \rho(W)$ and tuning $\rho(W)$ (or a scalar gain g in $W = gW_0$ with $\|W_0\| = 1$) directly tunes L: $L \approx (1-\alpha) + \alpha g \;\Rightarrow\; g \approx \frac{L-(1-\alpha)}{\alpha}$. This nonnegative-gain prescription requires $L \geq 1-\alpha$; otherwise the leak parameter α must be increased to make the target bound attainable. This gives a transparent leak–gain trade-off: increasing α makes the contraction more sensitive to g; for $\|W\| < 1$, increasing α decreases the Lipschitz bound $L = (1-\alpha) + \alpha\|W\|$ toward $\|W\|$ (stronger contraction, faster forgetting), whereas decreasing α increases L toward 1 (weaker contraction), increasing the dimension permitted by this uniform bound. **Rule 3: encourage separation by input-induced translations (to avoid dimension drop).** Exact formulas such as Theorem 4.5 require limited overlaps (OSC/SSC). In the similarity proxy $F_s(x) = rR_s x + b_s$, a simple sufficient condition for

SSC is that the image sets $F_s(K)$ are pairwise disjoint; a crude but practical proxy is to make the translation vectors $\{b_s\}$ well separated relative to the image diameter scale $r\,\mathrm{diam}(K)$ [14]. In ESNs, increasing input scaling can enlarge effective separation between input-conditioned images, but this must be balanced against nonlinearity saturation (which can reduce effective entropy transmission and increase contraction).

6 Implications for Learning with Linear Readouts

Reservoir computing trains only a linear readout on samples of the (dependent) reservoir state process. State geometry and temporal dependence provide complementary information about the representation. Contractivity quantifies Wasserstein convergence and Lipschitz-observable correlation decay, while the dimension bounds quantify geometric complexity. A direct implication from information dimension to linear-readout risk is not established here.

6.1 Mixing from contraction and consequences for generalization

Assume i.i.d. inputs $U_t \sim p$, so that (X_t) is a time-homogeneous Markov chain with transition kernel $P(x,\cdot) := \sum_{s\in\Sigma} p(s)\,\delta_{F_s(x)}(\cdot)$. Uniform contractivity implies a strong coupling property: if two copies of the chain are driven by the same input sequence but start from different initial conditions, their distance contracts deterministically at rate L^t. This yields geometric ergodicity in Wasserstein distance and exponential decay of Lipschitz correlations, standard for iterated random functions [9, 10].

Proposition 6.1 (Wasserstein mixing for contractive reservoirs)**.** *Assume Assumption 2.1 and i.i.d. inputs. Let $X_t^{(x)}$ denote the chain started at $X_0 = x$. Then for any $x, y \in X$ and any $t \geq 0$,*

$$W_1\Big(\mathcal{L}(X_t^{(x)}),\, \mathcal{L}(X_t^{(y)})\Big) \;\leq\; L^t \|x - y\|. \tag{7}$$

In particular, there is a unique stationary distribution μ, and

$$W_1\Big(\mathcal{L}(X_t^{(x)}),\, \mu\Big) \;\leq\; L^t\,\mathrm{diam}(X). \tag{8}$$

Consequence: exponential decay of correlations (Lipschitz observables). Let $f : X \to \mathbb{R}$ be bounded and $g : X \to \mathbb{R}$ be Lipschitz. Under stationarity, $|\mathrm{Cov}(f(X_0), g(X_t))| \leq 2\|f\|_\infty \mathrm{Lip}(g)\,\mathrm{diam}(X)\,L^t$, by bounding $|\mathbb{E}[g(X_t) \mid X_0] - \mathbb{E}[g(X_t)]|$ using (8). A proof for i.i.d. inputs is given in Appendix A.5; general stationary inputs require additional dependence assumptions.

For bounded Lipschitz observables, the summable covariance bound yields sample-mean variance $O([T(1-L)]^{-1})$, with constants depending on the observable and state space. General regression concentration and excess-risk guarantees require additional assumptions on the targets, noise, and dependence structure and are not established here.

6.2 Effective dimension and sample complexity

A linear readout trained by ridge regression uses the design matrix of reservoir states, so its statistical behavior depends on the state second-moment matrix $C := \mathbb{E}[X_0 X_0^\top]$ (the covariance for centered states) (under stationarity) and on the dependence across time. A standard complexity proxy is the *effective dimension* (a.k.a. effective degrees of freedom/statistical dimension) at regularization level $\lambda > 0$: $d_{\mathrm{eff}}(\lambda) := \mathrm{tr}\big(C(C+\lambda I)^{-1}\big)$, which governs the variance term in ridge regression and appears throughout learning theory for linear and kernel methods [23, 5].

For a well-specified linear model $Y_t = \langle w_\star, X_t\rangle + \xi_t$ with conditionally zero-mean noise of bounded variance, standard i.i.d. ridge-regression bounds involve a variance term of order $\sigma^2 d_{\mathrm{eff}}(\lambda)/T$, together with regularization bias and suitable design assumptions [23, 5]. Under appropriate mixing assumptions, dependence-dependent effective sample sizes enter analogous bounds [42, 34]. Proposition 6.1 controls Lipschitz observables; it does not by itself supply all assumptions for these regression results. An effective sample size can remain a contraction-dependent fraction of T even for long trajectories.

Geometric takeaway. Information dimension and covariance effective dimension measure different properties: even a finite-support distribution can have zero information dimension and full-rank

covariance. The entropy–contraction bound therefore complements, rather than replaces, task-specific memory and capacity measures [26, 32]. The experiments examine whether these quantities vary together in the reported settings. A rigorous link from information dimension to dependent-sample readout risk is left to future work.

Empirical Validation (see Appendix) Appendix B provides an empirical study of stationary reservoir geometry. We compare finite-sample dimension estimates using several estimators and time-series precautions [30, 13, 20]; calibrate against the exact similarity law under OSC [25, 14]; and examine entropy and contraction sweeps for nonlinear ESNs. Tangent Lyapunov estimates are empirical proxies, distinct from the Lipschitz exponent in Theorem 3.1. Small-scale readout experiments explore associations with memory and prediction, without establishing a universal performance law or a full-dimensional phase transition.

7 Discussion, Limitations and Conclusion

We introduced a partition-free, geometric theory of contractive reservoir dynamics: contractive ESNs with quantized inputs define a finite IFS/RDS [25, 1], whose attractor is the reachable state set K and whose stationary distribution μ captures the representation actually used by readouts. The entropy–contraction bound (Theorem 3.1) links the *information-theoretic complexity* of inputs (entropy rate) to the *geometric complexity* of reservoir states (intrinsic dimension), and the similarity/conformal regime yields exact closed forms under standard separation hypotheses (Theorems 4.4–4.5) [35, 14]. This identifies an explicit necessary entropy–contraction threshold for full-dimensionality inside the stable ESP regime. Attaining that threshold does not in general guarantee full dimension.

Limitations. *Quantized inputs.* Our IFS framing is exact for finite alphabets and for quantized inputs. For continuous inputs, one must either quantize with controlled error, or treat inputs as parameters in a continuous family of contractions and work with more general random dynamical systems; extending the entropy–dimension bounds to continuous alphabets is a natural direction. *Overlaps and dimension drop.* Exact formulas require separation conditions (OSC/SSC). Without separation, self-similar measures can exhibit dimension drop due to exact or approximate overlaps [14]. Our general bound remains valid without separation, but understanding overlap-induced reduction (and how ESN parameter choices control overlaps) is an important open direction. *Beyond conformal contraction.* Standard nonlinear ESNs need not be conformal; even their affine special cases can be anisotropic. Extending exact laws from similarity to general self-affine settings typically requires singular value pressure and subadditive thermodynamic formalism; our main-body focus is on robust bounds and a clean conformal reference model [14, 29]. *Near-critical regimes.* Our results apply in the strictly contractive regime. As $L \uparrow 1$, the guaranteed Wasserstein convergence and correlation-decay rates deteriorate, complicating finite-sample estimation; for $L \geq 1$, ESP may fail and the IFS/RDS picture changes qualitatively [26, 32]. Understanding how geometric complexity behaves near and beyond the stability boundary remains important for practice. *Learning-theoretic link.* We provided a principled connection between contraction (mixing) and representation geometry (dimension), and discussed implications for readouts via effective dimension. A full end-to-end theory connecting $\dim(\mu)$ to task risk for dependent samples is beyond the scope of this paper and is left to future work.

Empirical scope. The reported experiments use small reservoirs and few seeds. Finite-scale dimension estimates can be biased. These experiments are exploratory; generalization to larger or physical reservoirs remains to be assessed.

Broader perspective. Dimension is not a replacement for classical reservoir metrics (memory capacity, information processing capacity, approximation universality) but a complementary axis: it quantifies the size of the representation manifold/fractal explored by admissible inputs, while capacity metrics quantify what functions can be computed from those representations [8, 21]. We expect intrinsic dimension to be especially valuable as a design coordinate when stability, interpretability, or certification in the ESP regime matters.

Broader impacts. Better understanding of reservoir geometry may support more informed model design and resource allocation. A potential risk is treating dimension heuristics as guarantees of predictive accuracy or safety in downstream applications. Such applications require task-specific validation beyond the results proved here.

Conclusion. We developed a fractal-geometric view of contractive reservoirs: with quantized inputs, a contractive ESN induces an IFS/RDS whose attractor is the reachable state set and whose stationary measure describes the reservoir representation used by linear readouts [25, 1]. We proved entropy–contraction dimension bounds showing that intrinsic state dimension is controlled by input entropy rate divided by a contraction exponent (Theorem 3.1), and we identified a similarity/conformal regime where these bounds become exact closed-form laws under separation (Theorems 4.4–4.5) [14]. This yields a necessary entropy–contraction design criterion within the stable ESP regime, linking recurrent gain, leak, and input statistics to the dimension permitted by the bound. We hope this perspective helps turn reservoir design from heuristic tuning into principled control of representation geometry.

Acknowledgments and Disclosure of Funding

Funding. This research received no specific funding from any public, commercial, or not-for-profit funding agency.

Competing interests. The authors declare that they have no competing financial interests related to this work.

References

[1] L. Arnold. Random Dynamical Systems. Springer, 1998.

[2] M. F. Barnsley. Fractals Everywhere. Academic Press, Boston, 1988.

[3] N. Bertschinger and T. Natschläger. Real-time computation at the edge of chaos in recurrent neural networks. Neural Computation, 16(7):1413–1436, 2004.

[4] J. Boedecker, O. Obst, J. Lizier, N. Mayer, and M. Asada. Information processing in echo state networks at the edge of chaos. Theory in Biosciences, 131(3):205–213, 2012.

[5] A. Caponnetto and E. Vito. Optimal rates for the regularized least-squares algorithm. Found. Comput. Math., 7(3):331–368, July 2007.

[6] T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley, 2 edition, 2006.

[7] C. Cuchiero, L. Gonon, L. Grigoryeva, J.-P. Ortega, and J. Teichmann. Discrete-time signatures and randomness in reservoir computing. IEEE Transactions on Neural Networks and Learning Systems, 33(11):6321–6330, 2022.

[8] J. Dambre, D. Verstraeten, B. Schrauwen, and S. Massar. Information processing capacity of dynamical systems. Scientific Reports, 2:514, 2012.

[9] P. Diaconis and D. Freedman. Iterated random functions. SIAM Rev., 41(1):45–76, Mar. 1999.

[10] P. Doukhan. Mixing: Properties and Examples, volume 85 of Lecture Notes in Statistics. Springer, 1994.

[11] J. H. Elton. An ergodic theorem for iterated maps. Ergodic Theory and Dynamical Systems, 7(4):481–488, 1987.

[12] N. B. Erichson, O. Azencot, A. Queiruga, L. Hodgkinson, and M. W. Mahoney. Lipschitz recurrent neural networks. In International Conference on Learning Representations, 2021.

[13] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. Scientific Reports, 7:12140, 2017.

[14] K. Falconer. Fractal Geometry: Mathematical Foundations and Applications. John Wiley & Sons, Chichester, 2 edition, 2003.

[15] C. Gallicchio, A. Micheli, and L. Pedrelli. Design of deep echo state networks. Neural Networks, 108:33–47, 2018.

[16] D. J. Gauthier, E. Bollt, A. Griffith, and W. A. S. Barbosa. Next generation reservoir computing. Nature Communications, 12:5564, 2021.

[17] L. Gonon, L. Grigoryeva, and J.-P. Ortega. Risk bounds for reservoir computing. Journal of Machine Learning Research, 21(240):1–61, 2020.

[18] L. Gonon and J.-P. Ortega. Reservoir computing universality with stochastic inputs. IEEE Transactions on Neural Networks and Learning Systems, 31(1):100–112, 2020.

[19] L. Gonon and J.-P. Ortega. Fading memory echo state networks are universal. Neural Networks, 138:10–13, 2021.

[20] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. Physica D: Nonlinear Phenomena, 9(1):189–208, 1983.

[21] L. Grigoryeva and J.-P. Ortega. Echo state networks are universal. Neural Networks, 108:495–508, 2018.

[22] L. Grigoryeva and J.-P. Ortega. Differentiable reservoir computing. Journal of Machine Learning Research, 20(179):1–62, 2019.

[23] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, 2nd edition, 2009.

[24] H. Hewamalage, C. Bergmeir, and K. Bandara. Recurrent neural networks for time series forecasting: Current status and future directions. International Journal of Forecasting, 37(1):388–427, 2021.

[25] J. E. Hutchinson. Fractals and self-similarity. Indiana University Mathematics Journal, 30(5):713–747, 1981.

[26] H. Jaeger. The "echo state" approach to analysing and training recurrent neural networks. GMD Technical Report, 148:1–47, 2001.

[27] H. Jaeger. Adaptive nonlinear system identification with echo state networks. In Advances in Neural Information Processing Systems (NIPS), volume 15, pages 593–600, 2002.

[28] J. F. C. Kingman. Subadditive ergodic theory. Annals of Probability, 1(6):883–909, 1973.

[29] F. Ledrappier and L. Young. Dimension formula for random transformations. Communications in Mathematical Physics, 117(4):529–548, 1988.

[30] E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. In Proceedings of the 18th International Conference on Neural Information Processing Systems, NIPS'04, page 777–784, Cambridge, MA, USA, 2004. MIT Press.

[31] M. Lukoševičius. A practical guide to applying echo state networks. In Neural Networks: Tricks of the Trade, volume 7700 of Lecture Notes in Computer Science, pages 659–686. Springer, 2012.

[32] M. Lukoševičius and H. Jaeger. Reservoir computing approaches to recurrent neural network training. Computer Science Review, 3(3):127–149, 2009.

[33] J. Miller and M. Hardt. Stable recurrent models. In International Conference on Learning Representations, 2019.

[34] M. Mohri, A. Rostamizadeh, A. Talwalkar, and F. Bach. Foundations of Machine Learning. MIT Press, 2 edition, 2018.

[35] P. A. P. Moran. Additive functions of intervals and hausdorff measure. Mathematical Proceedings of the Cambridge Philosophical Society, 42(1):15–23, 1946.

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

[37] A. Rényi. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 547–561. University of California Press, 1961.

[38] A. Schief. Separation properties for self-similar sets. Proceedings of the American Mathematical Society, 122(1):111–115, 1994.

[39] C. E. Shannon. A mathematical theory of communication. SIGMOBILE Mob. Comput. Commun. Rev., 5(1):3–55, Jan. 2001.

[40] Y. Wu and S. Verdú. Rényi information dimension: Fundamental limits of almost lossless analog compression. IEEE Transactions on Information Theory, 56(8):3721–3748, 2010.

[41] L.-S. Young. Dimension, entropy and lyapunov exponents. Ergodic Theory and Dynamical Systems, 2(1):109–124, 1982.

[42] B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. The Annals of Probability, 22(1):94–116, 1994.

A Proofs and Technical Details

This appendix provides complete proofs of all formal results stated in the main body. We follow the notation of the manuscript.

A.1 Preliminaries and auxiliary lemmas

Lipschitz constants and compositions. For a map $G : X \to X$, denote by $\mathrm{Lip}(G)$ its global Lipschitz constant on X:

$$\mathrm{Lip}(G) := \sup_{x \neq y \in X} \frac{\|G(x) - G(y)\|}{\|x - y\|}.$$

If G, H are Lipschitz, then $\mathrm{Lip}(G \circ H) \leq \mathrm{Lip}(G)\mathrm{Lip}(H)$. We use $\mathrm{Lip}(G) = 0$ when X is a singleton, and $\log 0 = -\infty$.

Quantizer. Fix $\varepsilon > 0$ and let $Q_\varepsilon : \mathbb{R}^N \to \mathbb{Z}^N$ be the axis-aligned uniform quantizer mapping x to the index of the ε-cube containing x (e.g. $Q_\varepsilon(x) = (\lfloor x_1/\varepsilon \rfloor, \ldots, \lfloor x_N/\varepsilon \rfloor)$). For a set $A \subset \mathbb{R}^N$, write

$$\mathcal{Q}_\varepsilon(A) \;:=\; \{Q_\varepsilon(x) : x \in A\}, \qquad M_\varepsilon(A) \;:=\; |\mathcal{Q}_\varepsilon(A)|.$$

Lemma A.1 (How many ε-cubes can a small set intersect?)**.** *If $A \subset \mathbb{R}^N$ has $\mathrm{diam}(A) \leq \varepsilon$, then $M_\varepsilon(A) \leq 2^N$. More generally, for any A and any $\varepsilon > 0$,*

$$M_\varepsilon(A) \;\leq\; \left(1 + \left\lceil \frac{\mathrm{diam}(A)}{\varepsilon} \right\rceil\right)^N \;\leq\; \left(2 + \frac{\mathrm{diam}(A)}{\varepsilon}\right)^N. \tag{9}$$

Proof. If $\mathrm{diam}(A) \leq \varepsilon$, then along each coordinate axis A spans an interval of length at most ε, hence intersects at most 2 adjacent ε-intervals in that coordinate. Therefore it intersects at most 2^N cubes. The general bound follows by the same coordinate-wise counting: along each axis, the number of ε-intervals intersected is at most $1 + \lceil \mathrm{diam}(A)/\varepsilon \rceil$. □

Lemma A.2 (Entropy bound from finite support)**.** *If a discrete random variable Z takes values in a finite set of size at most M, then $H(Z) \leq \log M$.*

Proof. The entropy is maximized by the uniform distribution on a fixed finite support. □

Entropy rate facts. For any stationary process on a finite alphabet, the block entropy is subadditive: $H(U_{0:n+m-1}) \leq H(U_{0:n-1}) + H(U_{n:n+m-1}) = H(U_{0:n-1}) + H(U_{0:m-1})$, hence the limit $h(U) = \lim_{n\to\infty} \frac{1}{n} H(U_{0:n-1})$ exists by Fekete's lemma [6].

A.2 Proof of Theorem 2.2 (Echo state canonical solution)

We prove existence and uniqueness of the pullback limit (2) and the fading-memory estimate.

Proof of Theorem 2.2. Fix a deterministic bi-infinite input realization $(U_t)_{t\in\mathbb{Z}}$ (we suppress ω for readability) and fix $t \in \mathbb{Z}$.

Step 1: existence of the pullback limit. For $n \geq 1$ define

$$Z_t^{(n)}(x^\star) \;:=\; F_{U_{t-1}} \circ F_{U_{t-2}} \circ \cdots \circ F_{U_{t-n}}(x^\star) \;\in X.$$

We show $(Z_t^{(n)}(x^\star))_{n\geq 1}$ is Cauchy in the complete metric space $(X, \|\cdot\|)$. Let $m > n$. Write

$$Z_t^{(m)}(x^\star) = F_{U_{t-1}} \circ \cdots \circ F_{U_{t-n}}\Big(F_{U_{t-n-1}} \circ \cdots \circ F_{U_{t-m}}(x^\star)\Big), \quad Z_t^{(n)}(x^\star) = F_{U_{t-1}} \circ \cdots \circ F_{U_{t-n}}(x^\star).$$

By Assumption 2.1 and the composition Lipschitz inequality,

$$\|Z_t^{(m)}(x^\star) - Z_t^{(n)}(x^\star)\| \leq \mathrm{Lip}\Big(F_{U_{t-1}} \circ \cdots \circ F_{U_{t-n}}\Big) \cdot \Big\|F_{U_{t-n-1}} \circ \cdots \circ F_{U_{t-m}}(x^\star) - x^\star\Big\|.$$

The first factor is at most L^n. The second factor is bounded by $\mathrm{diam}(X)$ since both terms lie in X. Hence $\|Z_t^{(m)}(x^\star) - Z_t^{(n)}(x^\star)\| \leq L^n \mathrm{diam}(X)$. Therefore $(Z_t^{(n)}(x^\star))$ is Cauchy and converges to a limit in X; call it X_t.

Step 2: independence of $x^\star$. Let $x^\star, y^\star \in X$. Then

$$\|Z_t^{(n)}(x^\star) - Z_t^{(n)}(y^\star)\| \leq \mathrm{Lip}\Big(F_{U_{t-1}} \circ \cdots \circ F_{U_{t-n}}\Big) \cdot \|x^\star - y^\star\| \leq L^n \|x^\star - y^\star\| \to 0.$$

Thus $\lim_{n\to\infty} Z_t^{(n)}(x^\star)$ is independent of $x^\star$.

Step 3: (X_t) solves the recursion and is bounded. By continuity of each F_s and the definition of X_t as a limit,

$$X_{t+1} = \lim_{n\to\infty} F_{U_t} \circ F_{U_{t-1}} \circ \cdots \circ F_{U_{t-n}}(x^\star) = F_{U_t}\Big(\lim_{n\to\infty} F_{U_{t-1}} \circ \cdots \circ F_{U_{t-n}}(x^\star)\Big) = F_{U_t}(X_t).$$

Since $X_t \in X$ for all t, the solution is bounded.

Step 4: uniqueness among bounded bi-infinite solutions. Suppose $(Y_t)_{t\in\mathbb{Z}}$ is another bounded solution of (1) with $Y_t \in X$ for all t. For any $n \geq 1$, iterating the recursion backwards gives

$$X_t = F_{U_{t-1}} \circ \cdots \circ F_{U_{t-n}}(X_{t-n}), \qquad Y_t = F_{U_{t-1}} \circ \cdots \circ F_{U_{t-n}}(Y_{t-n}).$$

Hence

$$\|X_t - Y_t\| \leq \mathrm{Lip}\Big(F_{U_{t-1}} \circ \cdots \circ F_{U_{t-n}}\Big) \cdot \|X_{t-n} - Y_{t-n}\| \leq L^n \mathrm{diam}(X) \to 0,$$

so $X_t = Y_t$ for all t.

Step 5: fading memory estimate. Let two input histories agree on the last m symbols, i.e. $U_{t-m:t-1} = \widetilde{U}_{t-m:t-1}$. Let $X_t, \widetilde{X}_t$ be the corresponding canonical states. Then for any $n \geq m$, the last m maps in the pullback composition are identical, so

$$X_t = F_{U_{t-1}} \circ \cdots \circ F_{U_{t-m}}(X_{t-m}), \qquad \widetilde{X}_t = F_{U_{t-1}} \circ \cdots \circ F_{U_{t-m}}(\widetilde{X}_{t-m}).$$

Therefore

$$\|X_t - \widetilde{X}_t\| \leq \mathrm{Lip}\Big(F_{U_{t-1}} \circ \cdots \circ F_{U_{t-m}}\Big) \cdot \|X_{t-m} - \widetilde{X}_{t-m}\| \leq L^m \mathrm{diam}(X).$$

This matches the stated inequality. □

A.3 Proof of Theorem 2.3 and the reachable-set interpretation

Proof of Theorem 2.3. Let $\mathcal{K}(X)$ be the collection of nonempty compact subsets of X with the Hausdorff metric d_H. Define the Hutchinson operator $\mathcal{H}(A) = \bigcup_{s\in\Sigma} F_s(A)$.

Step 1: $\mathcal{H}$ is a contraction on $(\mathcal{K}(X), d_H)$. Fix $A, B \in \mathcal{K}(X)$. For any $s \in \Sigma$, the map F_s is L-Lipschitz, so $d_H(F_s(A), F_s(B)) \le L\, d_H(A, B)$ (standard property of Hausdorff distance under Lipschitz maps). Using that Hausdorff distance between finite unions is bounded by the maximum of component-wise distances,

$$d_H(\mathcal{H}(A), \mathcal{H}(B)) = d_H\Big(\bigcup_s F_s(A),\, \bigcup_s F_s(B)\Big) \le \max_{s\in\Sigma} d_H(F_s(A), F_s(B)) \le L\, d_H(A, B).$$

Step 2: Banach fixed point theorem. Since $(\mathcal{K}(X), d_H)$ is complete and $\mathcal{H}$ is a contraction with factor $L < 1$, Banach's theorem implies there exists a unique fixed point $K \in \mathcal{K}(X)$ satisfying $K = \mathcal{H}(K)$, and for any $A \in \mathcal{K}(X)$, $\mathcal{H}^n(A) \to K$ in d_H as $n \to \infty$ [25]. □

Reachable set as the pullback image. Define the pullback (coding) map $\Psi : \Sigma^{\mathbb{Z}_{<0}} \to X$ by

$$\Psi(u_{-1}, u_{-2}, \ldots) := \lim_{n\to\infty} F_{u_{-1}} \circ F_{u_{-2}} \circ \cdots \circ F_{u_{-n}}(x^\star),$$

which exists and is independent of $x^\star$ by Theorem 2.2 applied at $t = 0$.

Proposition A.3 (Attractor equals pullback states)**.** *Under Assumption 2.1,*

$$K \;=\; \Psi(\Sigma^{\mathbb{Z}_{<0}}).$$

In particular, K is exactly the set of states reachable at time 0 from some infinite past.

Proof. We show two inclusions.

(1) $\Psi(\Sigma^{\mathbb{Z}_{<0}}) \subseteq K$. Fix a past $u_{-1:-\infty}$ and define $x_n := F_{u_{-1}} \circ \cdots \circ F_{u_{-n}}(x^\star) \in \mathcal{H}^n(\{x^\star\}) \subseteq \mathcal{H}^n(X)$. Since $\mathcal{H}^n(X) \to K$ in Hausdorff metric, any limit point of (x_n) belongs to K; in particular $\Psi(u_{-1:-\infty}) = \lim_{n\to\infty} x_n \in K$.

(2) $K \subseteq \Psi(\Sigma^{\mathbb{Z}_{<0}})$. Forward invariance gives $\mathcal{H}^{n+1}(X) \subseteq \mathcal{H}^n(X)$, so the Hausdorff convergence to K implies

$$K = \bigcap_{n\ge1} \mathcal{H}^n(X) = \bigcap_{n\ge1} \bigcup_{w\in\Sigma^n} F_{w_{n-1}} \circ \cdots \circ F_{w_0}(X).$$

Fix $x \in K$. For each n, choose a word $w^{(n)} \in \Sigma^n$ and $y_n \in X$ with $x = F_{w_{n-1}^{(n)}} \circ \cdots \circ F_{w_0^{(n)}}(y_n)$. Extend this word to an infinite past $u^{(n)}$ with $u^{(n)}_{-1:-n} = w^{(n)}_{n-1:0}$, and put $z_n = \Psi(u^{(n)})$. The two states share their last n maps, hence

$$\|x - z_n\| \le L^n \mathrm{diam}(X) \longrightarrow 0.$$

Thus $K \subseteq \overline{\Psi(\Sigma^{\mathbb{Z}_{<0}})}$. The fading-memory estimate in Theorem 2.2 makes Ψ continuous in the product topology. Since the finite-alphabet past space $\Sigma^{\mathbb{Z}_{<0}}$ is compact, its image is compact and therefore closed. This removes the closure and proves the desired inclusion. □

A.4 Proof of Theorem 2.4 (stationary IFS measure)

We prove existence and uniqueness of the invariant measure for i.i.d. inputs by contraction in Wasserstein distance (the classical "iterated random functions" argument [11, 9]).

Lemma A.4 (Pushforward contracts Wasserstein-1)**.** *Let $F : X \to X$ be L_F-Lipschitz. Then for any probability measures ν, ν' on X,*

$$W_1(F_\#\nu,\; F_\#\nu') \;\le\; L_F\, W_1(\nu, \nu').$$

Proof. Using the Kantorovich–Rubinstein dual representation of W_1,

$$W_1(\nu, \nu') = \sup_{\mathrm{Lip}(f) \le 1} \int f \, d\nu - \int f \, d\nu'.$$

For any f with $\mathrm{Lip}(f) \le 1$, the composition $f \circ F$ has Lipschitz constant at most L_F. Therefore

$$\int f \, d(F_{\#}\nu) - \int f \, d(F_{\#}\nu') = \int (f \circ F) \, d\nu - \int (f \circ F) \, d\nu' \le L_F \, W_1(\nu, \nu'),$$

and taking the supremum over $\mathrm{Lip}(f) \le 1$ yields the claim. □

Proof of Theorem 2.4. Define $\mathcal{T}_p(\nu) = \sum_{s \in \Sigma} p(s)(F_s)_{\#}\nu$ on $\mathcal{P}(X)$.

Step 1: $\mathcal{T}_p$ is a contraction in W_1. By convexity of W_1 under mixtures and Lemma A.4,

$$\begin{aligned} W_1(\mathcal{T}_p(\nu), \mathcal{T}_p(\nu')) &\le \sum_{s \in \Sigma} p(s) \, W_1((F_s)_{\#}\nu, (F_s)_{\#}\nu') \\ &\le \sum_{s \in \Sigma} p(s) \, \mathrm{Lip}(F_s) \, W_1(\nu, \nu') \\ &\le L \, W_1(\nu, \nu'). \end{aligned}$$

Thus $\mathcal{T}_p$ is a strict contraction on the complete metric space $(\mathcal{P}(X), W_1)$.

Step 2: Banach fixed point theorem. By Banach's theorem, there exists a unique $\mu_p \in \mathcal{P}(X)$ with $\mu_p = \mathcal{T}_p(\mu_p)$, and for any initial ν, $W_1(\mathcal{T}_p^n(\nu), \mu_p) \le L^n W_1(\nu, \mu_p) \to 0$.

Step 3: stationarity of the Markov chain. Let $X_0 \sim \mu_p$ and $U_t \sim p$ i.i.d. independent of X_0, and define $X_{t+1} = F_{U_t}(X_t)$. Then

$$\mathcal{L}(X_1) = \sum_{s \in \Sigma} p(s) \, (F_s)_{\#}\mathcal{L}(X_0) = \mathcal{T}_p(\mu_p) = \mu_p,$$

and by induction $\mathcal{L}(X_t) = \mu_p$ for all t, so (X_t) is stationary.

Step 4: support inclusion $\mathrm{supp}(\mu_p) \subseteq K$. Let $S := \mathrm{supp}(\mu_p)$ (closed). From invariance $\mu_p = \sum_s p(s)(F_s)_{\#}\mu_p$ it follows that $S \subseteq \bigcup_s F_s(S)$: indeed, if $x \notin \bigcup_s F_s(S)$ then some neighborhood of x has zero mass under each $(F_s)_{\#}\mu_p$ and thus under μ_p. Therefore S is a nonempty compact set satisfying $S \subseteq \mathcal{H}(S)$. Monotonicity gives $S \subseteq \mathcal{H}^n(S)$ for every n, and $\mathcal{H}^n(S) \to K$ in Hausdorff distance by Theorem 2.3. Hence $S \subseteq K$. □

A.5 Proof of Proposition 6.1 and the correlation decay bound

Proof of Proposition 6.1. Let $X_t^{(x)}$ and $X_t^{(y)}$ be two copies of the chain driven by the same i.i.d. input sequence (U_t), with $X_0^{(x)} = x$ and $X_0^{(y)} = y$. Then by Assumption 2.1, for every $t \ge 0$,

$$\|X_{t+1}^{(x)} - X_{t+1}^{(y)}\| = \|F_{U_t}(X_t^{(x)}) - F_{U_t}(X_t^{(y)})\| \le L \|X_t^{(x)} - X_t^{(y)}\|.$$

By induction, $\|X_t^{(x)} - X_t^{(y)}\| \le L^t \|x - y\|$ almost surely. This defines a coupling of the laws of $X_t^{(x)}$ and $X_t^{(y)}$, hence

$$W_1(\mathcal{L}(X_t^{(x)}), \mathcal{L}(X_t^{(y)})) \le \mathbb{E}\|X_t^{(x)} - X_t^{(y)}\| \le L^t \|x - y\|.$$

For the convergence to stationarity, let μ be the unique invariant measure from Theorem 2.4. Then the Markov operator P satisfies $W_1(P\nu, P\nu') \le L W_1(\nu, \nu')$ (same argument as in Theorem 2.4), so

$$W_1(\mathcal{L}(X_t^{(x)}), \mu) = W_1(P^t \delta_x, P^t \mu) \le L^t W_1(\delta_x, \mu) \le L^t \, \mathrm{diam}(X).$$

□

Correlation decay (as stated in the main text). Let f be bounded and g Lipschitz. Under stationarity, write

$$\mathrm{Cov}(f(X_0), g(X_t)) = \mathbb{E}\Big[f(X_0)\big(\mathbb{E}[g(X_t) \mid X_0] - \mathbb{E}[g(X_t)]\big)\Big].$$

Hence

$$|\mathrm{Cov}(f(X_0), g(X_t))| \le \|f\|_\infty \, \mathbb{E}\Big|\mathbb{E}[g(X_t) \mid X_0] - \mathbb{E}[g(X_t)]\Big|.$$

For each $x \in X$,

$$|\mathbb{E}[g(X_t) \mid X_0 = x] - \mathbb{E}[g(X_t)]| \le \mathrm{Lip}(g)\, W_1(\mathcal{L}(X_t^{(x)}), \mu) \le \mathrm{Lip}(g)\,\mathrm{diam}(X)\, L^t,$$

using Kantorovich–Rubinstein duality and Proposition 6.1. Combining yields

$$|\mathrm{Cov}(f(X_0), g(X_t))| \le \|f\|_\infty \,\mathrm{Lip}(g)\,\mathrm{diam}(X)\, L^t,$$

and therefore also the slightly looser bound stated in the main text with a factor 2.

A.6 Proof of Theorem 3.1 (entropy–contraction dimension bound)

We prove both parts of Theorem 3.1 via a quantitative "conditioning shrinks diameter" argument and the definition of information dimension [37, 40]. If $\mathrm{diam}(X) = 0$, the stationary measure is a point mass and its information dimension is zero, so both bounds are immediate. Hence assume $\mathrm{diam}(X) > 0$ below.

Conditional state sets. Fix $t = 0$ for simplicity. For $n \ge 1$ and a past block $w = u_{-n:-1} \in \Sigma^n$, define the set of possible present states consistent with that block:

$$S_n(w) \;:=\; \Big\{\, X_0(u_{-\infty:-1}) \;:\; u_{-n:-1} = w \,\Big\} \;\subseteq X.$$

By Theorem 2.2, X_0 is well-defined for every past.

Lemma A.5 (Diameter of a cylinder-conditioned state set)**.** *Under Assumption 2.1, for every n and every $w \in \Sigma^n$,*

$$\mathrm{diam}(S_n(w)) \;\le\; L^n\,\mathrm{diam}(X). \tag{10}$$

More generally, let

$$\Lambda_n(w) \;:=\; \log \mathrm{Lip}\Big(F_{w_{-1}} \circ \cdots \circ F_{w_{-n}}\Big),$$

then

$$\mathrm{diam}(S_n(w)) \;\le\; e^{\Lambda_n(w)}\,\mathrm{diam}(X). \tag{11}$$

Proof. Take two pasts $u_{-\infty:-1}$ and $\tilde{u}_{-\infty:-1}$ that share the same last n symbols w. Let X_{-n} and $\tilde{X}_{-n}$ be their canonical states at time $-n$. Then

$$X_0 = F_{w_{-1}} \circ \cdots \circ F_{w_{-n}}(X_{-n}), \qquad \tilde{X}_0 = F_{w_{-1}} \circ \cdots \circ F_{w_{-n}}(\tilde{X}_{-n}).$$

Therefore

$$\|X_0 - \tilde{X}_0\| \le \mathrm{Lip}\Big(F_{w_{-1}} \circ \cdots \circ F_{w_{-n}}\Big) \cdot \|X_{-n} - \tilde{X}_{-n}\| \le e^{\Lambda_n(w)}\,\mathrm{diam}(X).$$

This proves (11); the uniform bound (10) follows since $\Lambda_n(w) \le n \log L$ by Assumption 2.1. □

Quantization entropy via conditioning. Let $Z := X_0$ and consider $Q_\varepsilon(Z)$.

Lemma A.6 (Conditioning bound for quantized entropy)**.** *For any $n \ge 1$ and any $\varepsilon > 0$,*

$$H(Q_\varepsilon(Z)) \;\le\; H(U_{-n:-1}) \;+\; \mathbb{E}\big[\log M_\varepsilon(S_n(U_{-n:-1}))\big]. \tag{12}$$

Proof. By the chain rule and nonnegativity of conditional entropy,

$$H(Q_\varepsilon(Z)) \le H(U_{-n:-1}, Q_\varepsilon(Z)) = H(U_{-n:-1}) + H(Q_\varepsilon(Z) \mid U_{-n:-1}).$$

Conditioning on $U_{-n:-1} = w$, the random variable Z takes values in $S_n(w)$, hence $Q_\varepsilon(Z)$ takes values in $\mathcal{Q}_\varepsilon(S_n(w))$ of size $M_\varepsilon(S_n(w))$. By Lemma A.2, $H(Q_\varepsilon(Z) \mid U_{-n:-1} = w) \le \log M_\varepsilon(S_n(w))$. Taking expectation over w yields (12). □

Proof of Theorem 3.1(1): uniform bound

Proof of Theorem 3.1(1). Fix $0 < \varepsilon < \operatorname{diam}(X)$ and choose

$$n(\varepsilon) \;:=\; \left\lceil \frac{\log(\operatorname{diam}(X)/\varepsilon)}{|\log L|} \right\rceil.$$

Then $L^{n(\varepsilon)}\operatorname{diam}(X) \le \varepsilon$. By Lemma A.5, for every block w of length $n(\varepsilon)$, $\operatorname{diam}(S_{n(\varepsilon)}(w)) \le \varepsilon$, hence by Lemma A.1, $M_\varepsilon(S_{n(\varepsilon)}(w)) \le 2^N$. Plugging into Lemma A.6 yields

$$H(Q_\varepsilon(Z)) \le H(U_{-n(\varepsilon):-1}) + N\log 2. \tag{13}$$

By stationarity, $H(U_{-n:-1}) = H(U_{0:n-1})$, and by definition of entropy rate,

$$H(U_{0:n-1}) = n\,h(U) + o(n) \qquad (n \to \infty).$$

Also, $\log(1/\varepsilon) = n(\varepsilon)|\log L| + O(1)$ as $\varepsilon \downarrow 0$. Divide (13) by $\log(1/\varepsilon)$ and take $\limsup_{\varepsilon\downarrow 0}$:

$$\overline{\dim}_I(\mu) = \limsup_{\varepsilon\downarrow 0} \frac{H(Q_\varepsilon(Z))}{\log(1/\varepsilon)} \le \limsup_{\varepsilon\downarrow 0} \frac{n(\varepsilon)\,h(U) + o(n(\varepsilon)) + O(1)}{n(\varepsilon)|\log L| + O(1)} = \frac{h(U)}{|\log L|}.$$

Finally, $\overline{\dim}_I(\mu) \le N$ always, since $Q_\varepsilon(Z)$ takes values in at most $O((\operatorname{diam}(X)/\varepsilon)^N)$ bins (Lemma A.1), hence $H(Q_\varepsilon(Z)) \le N\log(1/\varepsilon) + O(1)$. Thus $\overline{\dim}_I(\mu) \le \min\{N, h(U)/|\log L|\}$. □

Proof of Theorem 3.1(2): refined pathwise bound

We now show how to replace $|\log L|$ by a (typically larger) effective contraction rate $|\chi|$, where χ is the subadditive contraction exponent defined in the main text.

Shift setup and contraction exponent. Let $\Omega = \Sigma^{\mathbb{Z}}$ with its product σ-algebra and stationary measure $\mathbb{P}$ induced by the input process. Let $\sigma : \Omega \to \Omega$ be the left shift $(\sigma u)_t = u_{t+1}$. Define

$$\Lambda_n(u) \;:=\; \log \operatorname{Lip}\Big(F_{u_{n-1}} \circ \cdots \circ F_{u_0}\Big).$$

Then subadditivity holds:

$$\begin{aligned}
\Lambda_{n+m}(u) &= \log \operatorname{Lip}\Big(F_{u_{n+m-1}} \circ \cdots \circ F_{u_0}\Big) \\
&\le \log \operatorname{Lip}\Big(F_{u_{n+m-1}} \circ \cdots \circ F_{u_m}\Big) \\
&\quad + \log \operatorname{Lip}\Big(F_{u_{m-1}} \circ \cdots \circ F_{u_0}\Big) \\
&= \Lambda_n(\sigma^m u) + \Lambda_m(u).
\end{aligned}$$

For the refined bound we use the additional hypothesis in Theorem 3.1(ii): $\Lambda_n/n \to \chi$ almost surely for a deterministic constant $\chi \in (-\infty, 0)$. Uniform contractivity gives $\Lambda_n/n \le \log L$, hence $\chi \le \log L < 0$. The argument below uses bounded positive parts and does not require L^1 convergence of Λ_n/n.

Proof of Theorem 3.1(2). Fix $0 < \varepsilon < \operatorname{diam}(X)$ and choose

$$n(\varepsilon) \;:=\; \left\lceil \frac{\log(\operatorname{diam}(X)/\varepsilon)}{|\chi|} \right\rceil.$$

Let $Z = X_0$ as before and condition on the length-$n(\varepsilon)$ past block $U_{-n(\varepsilon):-1}$. For this block w, Lemma A.5 gives $\operatorname{diam}(S_{n(\varepsilon)}(w)) \le \operatorname{diam}(X)\, e^{\Lambda_{n(\varepsilon)}(w)}$. Lemma A.1 then yields

$$\begin{aligned}
\log M_\varepsilon(S_{n(\varepsilon)}(w)) &\le N \log\left(2 + \frac{\operatorname{diam}(X) e^{\Lambda_{n(\varepsilon)}(w)}}{\varepsilon}\right) \\
&\le N\log 3 + N\Big(\log \frac{\operatorname{diam}(X)}{\varepsilon} + \Lambda_{n(\varepsilon)}(w)\Big)_+,
\end{aligned}$$

using $\log(2+e^a) \le \log 3 + a_+$. By stationarity, the log-Lipschitz constant of a length-n past block has the same distribution as the forward-block variable Λ_n. Taking expectations in Lemma A.6, we may therefore write

$$H(Q_\varepsilon(Z)) \le H(U_{-n(\varepsilon):-1}) + N\log 3 + N\,\mathbb{E}\Big[\Big(\log\frac{\operatorname{diam}(X)}{\varepsilon} + \Lambda_{n(\varepsilon)}\Big)_+\Big]. \tag{14}$$

By the choice of $n(\varepsilon)$, $\log(\operatorname{diam}(X)/\varepsilon) + n(\varepsilon)\chi \le 0$, so

$$\Big(\log\frac{\operatorname{diam}(X)}{\varepsilon} + \Lambda_{n(\varepsilon)}\Big)_+ \le n(\varepsilon)\Big(\frac{\Lambda_{n(\varepsilon)}}{n(\varepsilon)} - \chi\Big)_+.$$

Moreover, uniform contractivity gives the deterministic bound

$$0 \le \Big(\frac{\Lambda_n}{n} - \chi\Big)_+ \le (\log L - \chi)_+.$$

The assumed almost-sure convergence and bounded convergence imply $\mathbb{E}[(\Lambda_n/n - \chi)_+] \to 0$. Therefore the expectation term in (14) is $o(n(\varepsilon))$.

Finally, as in the uniform case, $H(U_{-n(\varepsilon):-1}) = n(\varepsilon)h(U) + o(n(\varepsilon))$, and $\log(1/\varepsilon) = n(\varepsilon)|\chi| + O(1)$. Divide (14) by $\log(1/\varepsilon)$ and take $\limsup_{\varepsilon\downarrow 0}$ to get

$$\overline{\dim}_I(\mu) \le \frac{h(U)}{|\chi|}.$$

Adding the trivial bound $\overline{\dim}_I(\mu) \le N$ yields the stated $\min\{N, h(U)/|\chi|\}$. □

Remark (reachable-set Minkowski bound). The covering argument for K stated in the main text is also immediate from IFS structure:

Proposition A.7 (Minkowski bound for the attractor K)**.** *Under Assumption 2.1,*

$$\overline{\dim}_M(K) \;\le\; \min\left\{N,\; \frac{\log|\Sigma|}{|\log L|}\right\}.$$

Proof. For each word $w \in \Sigma^n$, the set $F_{w_{n-1}} \circ \cdots \circ F_{w_0}(X)$ has diameter at most $L^n\operatorname{diam}(X)$. Since $K = \mathcal{H}^n(K) \subseteq \bigcup_{w\in\Sigma^n} F_w(X)$, we can cover K by at most $|\Sigma|^n$ sets of diameter $\le L^n\operatorname{diam}(X)$. Thus $N_{cL^n}(K) \le |\Sigma|^n$ for an appropriate constant c. Taking $\limsup$ yields $\overline{\dim}_M(K) \le \log|\Sigma|/|\log L|$, and $\overline{\dim}_M(K) \le N$ always. □

A.7 Proofs of Theorems 4.4 and 4.5 (similarity regime)

Throughout this subsection we assume Assumption 4.1: $F_s(x) = rR_sx + b_s$ with a common contraction ratio $r \in (0,1)$, and the open set condition (OSC) from Definition 4.2. Let K be the self-similar attractor and let μ be the self-similar measure associated with weights p. The case $|\Sigma| = 1$ gives a singleton attractor and a point mass, so both dimension formulas hold with value zero; assume $|\Sigma| \ge 2$ below.

Exact dimension of the set K

Proof of Theorem 4.4. Let $s_{\mathrm{sim}} = \frac{\log|\Sigma|}{|\log r|}$.

Upper Minkowski bound. As in Proposition A.7 with $L = r$,

$$N_{cr^n}(K) \;\le\; |\Sigma|^n \quad\Rightarrow\quad \overline{\dim}_M(K) \;\le\; \frac{\log|\Sigma|}{|\log r|} = s_{\mathrm{sim}}.$$

Thus $\dim_H(K) \le \overline{\dim}_M(K) \le s_{\mathrm{sim}}$.

Lower Hausdorff bound via a Frostman measure. Consider the uniform self-similar measure ν on K given by weights $p(s) = 1/|\Sigma|$. Equivalently, let $(W_k)_{k\geq 1}$ be i.i.d. uniform on Σ and define the coding map

$$\pi(W) \;:=\; \lim_{n\to\infty} F_{W_1} \circ \cdots \circ F_{W_n}(x^\star) \;\in K; \qquad \nu := \mathcal{L}(\pi(W)).$$

Then for each cylinder word $w \in \Sigma^n$, the level-n cylinder set $K_w := F_{w_1} \circ \cdots \circ F_{w_n}(K)$ has diameter $\asymp r^n$ and $\nu(K_w) = |\Sigma|^{-n} = r^{ns_{\mathrm{sim}}}$.

Under OSC, different cylinder images of a suitable open set are disjoint, implying in particular a bounded overlap property: there exists a constant C_{ov} (depending on the OSC open set and N) such that any ball of radius comparable to r^n intersects at most C_{ov} cylinder sets K_w of level n [14, 38]. Hence for any $x \in K$ and any ρ with $r^{n+1} < \rho \leq r^n$,

$$\nu(B(x,\rho)) \leq C_{\mathrm{ov}} \cdot r^{ns_{\mathrm{sim}}} \leq C\, \rho^{s_{\mathrm{sim}}}$$

for a constant C independent of x, ρ. By Frostman's lemma (mass distribution principle), this implies $\underline{\dim}_H(K) \geq s_{\mathrm{sim}}$ [14]. Therefore $\dim_H(K) = s_{\mathrm{sim}}$ and since $\overline{\dim}_M(K) \leq s_{\mathrm{sim}}$, we have equality $\overline{\dim}_M(K) = \dim_H(K) = s_{\mathrm{sim}}$.

On the $\min\{N, \cdot\}$ truncation. For similarity IFS under OSC, the volume packing constraint implies $|\Sigma| r^N \leq 1$, hence $s_{\mathrm{sim}} \leq N$. Thus the $\min\{N, s_{\mathrm{sim}}\}$ truncation in the main statement is harmless (and redundant under OSC). □

Exact dimension of the self-similar measure μ

Proof of Theorem 4.5. We may discard symbols with zero probability and, for this proof, take K to be the attractor of the remaining subsystem. This preserves OSC, the measure μ, and $H(p)$. If only one symbol remains, μ is a point mass and the conclusion is immediate; otherwise all retained weights are positive. Let $(W_k)_{k\geq 1}$ be i.i.d. with $\mathbb{P}(W_k = s) = p(s)$ and define the coding map π as above. Then $\mu = \mathcal{L}(\pi(W))$, and for any word $w = w_1 \cdots w_n \in \Sigma^n$,

$$\mu(K_w) = p(w_1) \cdots p(w_n), \qquad \operatorname{diam}(K_w) \asymp r^n.$$

Step 1: typical cylinder masses. By the Shannon–McMillan theorem for i.i.d. sources (or simply the law of large numbers), for $\mathbb{P}$-a.e. W,

$$-\frac{1}{n} \log \mu(K_{W_1 \cdots W_n}) = -\frac{1}{n} \sum_{k=1}^{n} \log p(W_k) \longrightarrow \sum_{s\in\Sigma} p(s)\,(-\log p(s)) = H(p).$$

Step 2: ball–cylinder comparability under OSC. Under OSC, there exist constants $c_1, c_2 > 0$ and a uniform overlap constant C_{ov} such that for μ-a.e. $x = \pi(W)$ and for all large n,

$$K_{W_1 \cdots W_n} \;\subseteq\; B(x, c_2 r^n),$$

and

$$B(x, c_1 r^n) \cap K \;\subseteq \bigcup_{w \in \mathcal{N}_n(W)} K_w, \qquad |\mathcal{N}_n(W)| \leq C_{\mathrm{ov}},$$

where $\mathcal{N}_n(W)$ is a set of at most C_{ov} words of length n indexing cylinders that intersect $B(x, c_1 r^n)$ [14]. Consequently,

$$\mu(B(x, c_2 r^n)) \;\geq\; \mu(K_{W_1 \cdots W_n}) = p(W_1) \cdots p(W_n),$$

and

$$\begin{aligned} \mu(B(x, c_1 r^n)) &\leq \sum_{w \in \mathcal{N}_n(W)} \mu(K_w) \\ &\leq C_{\mathrm{ov}} \max_{w \in \mathcal{N}_n(W)} \mu(K_w) \\ &\leq C_{\mathrm{ov}}\, \mu(K_{W_1 \cdots W_n})\, e^{o(n)}, \end{aligned}$$

The last inequality follows from a summable probability bound. Write $p_w := \prod_{i=1}^{n} p(w_i)$ and $W_{1:n} := W_1 \cdots W_n$. For words of length n, declare $v \sim w$ when $\operatorname{dist}(K_v, K_w) \leq c_1 r^n$. The OSC

packing argument above, applied to a ball containing K_w and its $c_1 r^n$-neighborhood, bounds the number of neighbors of each word by a constant M independent of n. For every $\delta > 0$,

$$\begin{aligned}\mathbb{P}\big(\exists v \sim W_{1:n} : p_v > e^{\delta n} p_{W_{1:n}}\big) &\le \sum_{\substack{w,v\in\Sigma^n:\ v\sim w \\ p_v > e^{\delta n} p_w}} p_w \\ &\le e^{-\delta n} \sum_{v\in\Sigma^n} p_v \,\#\{w : w \sim v\} \\ &\le M e^{-\delta n}.\end{aligned}$$

Borel–Cantelli, followed by $\delta \downarrow 0$ through positive rationals, yields $\max_{v\sim W_{1:n}} p_v \le p_{W_{1:n}} e^{o(n)}$ almost surely. Since every word in $\mathcal{N}_n(W)$ is a neighbor of $W_{1:n}$, this proves the required comparison.

Step 3: local dimension computation. Let $\rho_n := r^n$. The preceding bounds at radii $c_1 r^n$ and $c_2 r^n$, Step 1, and monotonicity (shifting n by a fixed integer when necessary) give, for every fixed $c > 0$,

$$\frac{\log \mu(B(x, c\rho_n))}{\log \rho_n} = \frac{\log \mu(B(x, cr^n))}{n \log r} \longrightarrow \frac{H(p)}{|\log r|},$$

for μ-a.e. x, since $\log r < 0$. Because successive radii have fixed ratio r, monotonicity extends this limit to all radii tending to zero. Thus μ is exact-dimensional with local dimension $H(p)/|\log r|$. In particular,

$$\dim_H(\mu) = \dim_I(\mu) = \frac{H(p)}{|\log r|},$$

where the equality $\dim_I(\mu) = \dim_H(\mu)$ for compactly supported exact-dimensional measures is standard [14, 29]. As in the set case, OSC implies $H(p) \le \log|\Sigma| \le N|\log r|$, so the value is automatically $\le N$. This proves Theorem 4.5. □

Discussion. The similarity/conformal regime is included in the manuscript as a *reference model* where the entropy–contraction upper bound becomes an equality under a classical non-overlap hypothesis. In general ESNs, anisotropy (self-affine behavior) and overlaps can strictly reduce the realized dimension; the main-body theorems remain valid as upper bounds without any separation assumptions.

B Experiments

Our theory makes quantitative predictions about the *intrinsic geometric complexity* of reservoir state representations, quantified by the dimension of the stationary state distribution μ (Section 2.3). The experimental questions are: (i) can we reliably estimate intrinsic dimension of reservoir states from finite dependent trajectories; (ii) how do information-dimension estimates compare with the entropy–contraction bounds of Theorem 3.1 and, for i.i.d. similarity reservoirs under OSC, the exact law of Theorem 4.5; and (iii) how does downstream task performance vary with an empirical tangent-exponent dimension proxy in an exploratory study (Section 5)?

Conventions (logs and units). All entropies and logarithms in this appendix use natural logarithms (nats). Since our predictions depend on ratios of entropies and logarithms (e.g. $H(p)/|\log r|$), the resulting dimension is invariant to the log base as long as it is used consistently.

B.1 Experimental setup

Implementation and randomness. All experiments use double precision. Randomness enters through (i) reservoir initialization (random matrices) and (ii) input sequence generation. Unless stated otherwise, we average results across seeds and report mean ± standard deviation (s.d.) across seeds.

The reported protocols do not establish identical reservoir realizations across Tables 5, 6, and 8. We therefore interpret each sweep separately and do not treat differences between their nominally matching baselines as controlled comparisons. The reported trends are exploratory.

Family A: similarity reservoirs (sanity-check model with OSC). We instantiate the conformal proxy $F_s(x) = rR_sx + b_s,\ r \in (0,1),\ R_s \in O(N)$, driven by a finite alphabet $\Sigma = \{1, \dots, |\Sigma|\}$. For calibration we use a particularly clean OSC construction: state dimension: $N = 5$, contraction: $r = 1/2$, rotations: $R_s \equiv I$ (identity; still orthogonal), alphabet size: $|\Sigma| = 8$, and translations: we embed the 3-bit cube into the first three coordinates and keep the last two coordinates fixed: $b_s \in \{0, 1/2\}^3 \times \{0\}^2,\ s \in \Sigma$, i.e. the 8 translations are the corners of a 3D cube of sidelength $1/2$ inside $\mathbb{R}^5$.

With $r = 1/2$ and this choice of $\{b_s\}$, the sets $F_s((0,1)^3 \times (-\varepsilon, \varepsilon)^2)$ are pairwise disjoint for any fixed $\varepsilon \in (0,1)$, hence OSC holds (Definition 4.2). The sets $F_s(K)$ meet on boundaries, so SSC does not hold. In this regime, under i.i.d. inputs the exact formula $\dim(\mu) = \min\left\{N, \frac{H(p)}{|\log r|}\right\}$ holds (Theorem 4.5) [25, 14]. For the Markov sweeps, $h(U)/|\log r|$ is the upper-bound benchmark from Theorem 3.1; the exact law of Theorem 4.5 is invoked here only for i.i.d. inputs.

Family B: contractive ESNs (robustness model). We use leaky ESNs with $\tanh$ nonlinearity $x_{t+1} = (1-\alpha)x_t + \alpha \tanh(Wx_t + W_{\text{in}}s_t + b)$, where $s_t \in \mathbb{R}^d$ is the input symbol vector. Throughout the ESN dimension-validation sweeps we set $\alpha = 1$ and $b = 0$, so $x_{t+1} = \tanh(Wx_t + W_{\text{in}}s_t)$. Contractivity is enforced by scaling W so that its operator norm satisfies $\|W\| \approx L < 1$; since $\tanh$ is 1-Lipschitz, the global Lipschitz constant is $\leq \|W\|$.

Matrix generation and scaling. For a given reservoir size N:

- Draw W_0 with i.i.d. entries $\sim \mathcal{N}(0, 1/N)$, then estimate the spectral norm $\|W_0\|_2$ (largest singular value) by power iteration on $W_0^\top W_0$ (100 steps), and set $W = L_{\text{target}} \cdot W_0/\|W_0\|_2$.
- Draw W_{in} with i.i.d. entries $\sim \mathcal{N}(0, 1/d)$ and scale by an input gain (here fixed to 1 unless stated).

Input alphabets used. We use two concrete finite alphabets:

- **(ESN geometry sweeps)** $d = 5$, $\Sigma = \{\pm e_i\}_{i=1}^5$ (so $|\Sigma| = 10$), where e_i are standard basis vectors in $\mathbb{R}^5$.
- **(Tasks)** $d = |\Sigma| = 10$, $\Sigma = \{e_1, \dots, e_{10}\}$ (one-hot encoding of a 10-ary alphabet).

Input processes and entropy control. We use two stationary processes on the same finite alphabet Σ:

- **i.i.d. inputs:** $s_t \overset{\text{i.i.d.}}{\sim} p$. The entropy rate is $h(U) = H(p)$ [6].
- **Markov inputs:** a Markov chain with transition matrix P and stationary law π. The entropy rate is $h(U) = -\sum_i \pi_i \sum_j P_{ij} \log P_{ij}$.

To sweep entropy at fixed alphabet size, we use a one-heavy family $p_1 = q,\ p_{s\neq 1} = \frac{1-q}{|\Sigma|-1}$, so $H(p)$ decreases continuously as $q \to 1$.

For Markov inputs we use a *sticky-mixture* family on Σ: $P_{ii} = a + \frac{1-a}{|\Sigma|},\ P_{ij} = \frac{1-a}{|\Sigma|}\ (i \neq j)$, equivalently: with probability a keep the previous symbol, and with probability $1-a$ resample uniformly from Σ. This chain is doubly stochastic, hence has uniform stationary distribution $\pi_i = 1/|\Sigma|$, and a closed-form entropy rate $h(U) = -\Big(p_{\text{diag}} \log p_{\text{diag}} + (|\Sigma|-1)p_{\text{off}} \log p_{\text{off}}\Big),\ p_{\text{diag}} = a + \frac{1-a}{|\Sigma|},\ p_{\text{off}} = \frac{1-a}{|\Sigma|}$. Increasing a decreases $h(U)$.

Trajectories, burn-in, and thinning (dependence control). All state sequences are generated by a single long run and then thinned to reduce short-lag dependence in distance-based estimators.

- Similarity reservoirs: burn-in $T_{\text{burn}} = 2000$. Thinning lag $\tau = 10$ chosen so that $r^\tau = (1/2)^{10} \approx 9.8 \times 10^{-4}$.
- ESN geometry sweeps: burn-in $T_{\text{burn}} = 3000$. Thinning lag τ chosen as the smallest integer such that $L^\tau \leq 0.05$.

- Task runs: burn-in $T_{\mathrm{burn}} = 500$ (for supervised learning). For dimension estimation on task trajectories only (not for training), we use a milder lag τ chosen as the smallest integer such that $L^\tau \leq 0.5$ to maintain adequate sample size.

For Markov inputs, L^τ controls forgetting under a common drive, not the dependence of the input chain; thinning is an empirical precaution and does not establish independence.

B.2 Estimating intrinsic dimension from finite trajectories

The theory concerns the dimension of the stationary state distribution μ (e.g. information dimension $\dim_I(\mu)$; Section 2.3). We estimate dimension from a thinned sample $\{x_i\}_{i=1}^n$ using complementary estimators.

Preprocessing and numerics. Before any estimator, we (i) drop coordinates with empirical variance $< 10^{-12}$ (this matters for the similarity construction where two coordinates are identically zero), and (ii) standardize the remaining coordinates to mean 0 and variance 1 (computed on the sample). All distances are Euclidean in the standardized space.

(E1) Quantized-entropy (information-dimension) slope. To align directly with information dimension, we use coarse-grained entropy scaling $H_\varepsilon = -\sum_c \hat{p}_c(\varepsilon) \log \hat{p}_c(\varepsilon)$, where $\hat{p}_c(\varepsilon)$ is the empirical mass in axis-aligned ε-grid cells c of $\mathbb{R}^N$ (in standardized coordinates; the grid is anchored by the sample minimum in each coordinate). We estimate the slope of H_ε versus $\log(1/\varepsilon)$ by least squares over scales where occupancy is neither degenerate nor saturated. Concretely, we use the fixed scale list $\varepsilon \in \{0.5,\ 0.25,\ 0.125,\ 0.0625,\ 0.03125,\ 0.015625,\ 0.0078125,\ 0.00390625\}$ and fit using those ε for which the number of occupied cells lies in $[20,\ n/2]$.

(E2) TwoNN (scale-free check). We use the TwoNN estimator of [13] based on the ratio of the two nearest-neighbor distances. Let $r_{i,1}$ and $r_{i,2}$ be the distances from x_i to its first and second nearest neighbors (excluding itself), and define $\mu_i = r_{i,2}/r_{i,1}$. The TwoNN intrinsic-dimension estimate is $\widehat{d}_{\mathrm{TwoNN}} = \left(\frac{1}{n}\sum_{i=1}^n \log \mu_i\right)^{-1}$, which is a maximum-likelihood estimator under a locally homogeneous Poisson model and is used here as an empirical diagnostic [13].

(E3) Levina–Bickel kNN MLE (local dimension). We use the MLE estimator of Levina and Bickel [30] with a fixed $k = 20$: $\widehat{d}_{\mathrm{LB}}(x_i) = \left[\frac{1}{k-1}\sum_{j=1}^{k-1} \log \frac{T_k(x_i)}{T_j(x_i)}\right]^{-1}$, $\widehat{d}_{\mathrm{LB}} = \frac{1}{n}\sum_{i=1}^n \widehat{d}_{\mathrm{LB}}(x_i)$.

(E4) Correlation dimension (fractal sanity check). We compute $C(\varepsilon) = \frac{2}{n(n-1)}\sum_{i<j} \mathbf{1}\{\|x_i - x_j\| < \varepsilon\}$ [20]. We subsample $n_{\mathrm{sub}} = 3000$ points for similarity runs and $n_{\mathrm{sub}} = 500$ for ESN sweeps, then: (i) choose 25 logarithmically spaced ε values between the 5th and 50th percentiles of pairwise distances, (ii) compute $\log C(\varepsilon)$ and fit the slope of $\log C(\varepsilon)$ vs. $\log \varepsilon$ using only points where $C(\varepsilon) \in [0.02, 0.2]$ (or all points if fewer than 5 satisfy this). This estimates correlation dimension D_2, which need not equal information dimension. For an i.i.d. common-ratio self-similar measure under OSC, its benchmark is $D_2 = -\log(\sum_s p(s)^2)/|\log r|$.

B.2.1 Sensitivity to thinning and kNN hyperparameters

To verify that our intrinsic-dimension conclusions are not artifacts of a particular estimator setting, we perform a sensitivity check on a representative contractive ESN configuration (the *orthogonal* case of Table 7). We vary (i) the thinning lag τ used to reduce temporal dependence, and (ii) the neighborhood size k used by the Levina–Bickel kNN MLE. All estimators drop near-constant coordinates (if any) and standardize coordinates before computing distances. Table 1 shows that both TwoNN and Levina–Bickel estimates are stable across a $4\times$ range of thinning lags and across $k \in \{10, 20, 40\}$. This indicates relative stability for the configuration and settings tested; it does not establish estimator consistency or independence of the thinned samples.

Table 1: Sensitivity of kNN dimension estimators to thinning τ and neighborhood size k. Setup: same ESN as in the orthogonal case of Table 7 ($N = 30$, $L_{\text{target}} = 0.92$, uniform i.i.d. inputs on $|\Sigma| = 10$), but using $n = 1000$ samples per run. For each seed we simulate one long trajectory, then subsample it with different thinning lags $\tau \in \{18, 36, 72\}$. We report TwoNN and Levina–Bickel MLE with $k \in \{10, 20, 40\}$ (mean $\pm$ s.d. across 3 seeds).

Thinning τ	$\hat{d}_{\text{TwoNN}}$	$\hat{d}_{\text{LB}}(k = 10)$	$\hat{d}_{\text{LB}}(k = 20)$	$\hat{d}_{\text{LB}}(k = 40)$
18	7.28 ± 0.31	8.27 ± 0.37	7.74 ± 0.36	7.66 ± 0.32
36	6.84 ± 0.21	8.26 ± 0.31	7.75 ± 0.34	7.60 ± 0.30
72	7.16 ± 0.48	8.27 ± 0.43	7.70 ± 0.36	7.61 ± 0.31

B.3 Validating entropy–contraction predictions

Theorem 3.1 gives an information-dimension upper bound

$$d_{\text{pred}} = \min\left\{N,\ \frac{h(U)}{|\chi|}\right\}, \tag{15}$$

where $h(U)$ is the entropy rate of the input and $\chi < 0$ is the uniform or Lipschitz-based pathwise contraction exponent specified in that theorem. We compare three quantities:

- the *exact* law (similarity + OSC under i.i.d. inputs): $d_{\text{theory}} = \min\{N, H(p)/|\log r|\}$ (Theorem 4.5);
- a *conservative* ESN exponent $\chi = \log \|W\|$ (uniform Lipschitz bound);
- a *tangent Lyapunov* exponent $\hat{\chi}_{\tan}$ estimated from the Jacobian along a typical trajectory (Benettin method). This quantity is generally not identical to the Lipschitz-based pathwise contraction exponent χ appearing in Theorem 3.1; we use it only in the empirical proxy $d_{\text{pred}}(\hat{\chi}_{\tan}) = \min\{N, h(U)/|\hat{\chi}_{\tan}|\}$. This substitution is not a certified upper bound from Theorem 3.1.

Effective covariance dimension (second-order baseline). Alongside intrinsic-dimension estimators, we sometimes report a covariance-based *effective dimension* $d_{\text{eff}}(\lambda_0) = \operatorname{tr}\big(C(C + \lambda_0 I)^{-1}\big) = \sum_{j=1}^{N} \frac{\lambda_j}{\lambda_j + \lambda_0}$, where C is the empirical covariance of the sampled states (after centering), $\{\lambda_j\}$ are its eigenvalues, and $\lambda_0 = 10^{-3}$ is a fixed scale. This is an effective-rank statistic and can remain large even when local/fractal intrinsic dimension is small (e.g. in clustered or multi-modal regimes).

Lyapunov-based exponent estimate for ESNs. For $\alpha = 1$ and $b = 0$, the ESN update is $x_{t+1} = \tanh(Wx_t + W_{\text{in}}s_t)$, with Jacobian $J_t = \operatorname{diag}\big(1 - \tanh^2(Wx_t + W_{\text{in}}s_t)\big)\, W$. We estimate the top Lyapunov exponent using the standard Benettin renormalization procedure: initialize a random unit vector v_0, iterate $v_{t+1} = J_t v_t / \|J_t v_t\|$, and accumulate $\log \|J_t v_t\|$. We use burn-in $T_{\text{Lyap,burn}} = 1000$ and averaging window $T_{\text{Lyap}} = 2000$ for ESN sweeps (and $T_{\text{Lyap,burn}} = 200$, $T_{\text{Lyap}} = 800$ for task runs due to shorter sequences). We report $\hat{\chi}_{\tan} = \frac{1}{T}\sum_{t=1}^{T} \log \|J_t v_t\|$.

B.3.1 Similarity reservoirs: i.i.d. entropy sweep (exact calibration)

We generate i.i.d. symbols on $|\Sigma| = 8$ with the one-heavy distribution $p_1 = q$, $p_{s\neq 1} = (1 - q)/7$ and sweep $q \in \left\{\frac{1}{8},\ 0.30,\ 0.60,\ 0.90\right\}$. For each q we generate a trajectory of length $T = T_{\text{burn}} + n\tau$ with $T_{\text{burn}} = 2000$, $\tau = 10$, and collect $n = 10000$ thinned samples. Table 2 compares the exact prediction $d_{\text{theory}} = H(p)/|\log r|$ to dimension estimates from Section B.2.

Takeaway. The estimates decrease as entropy decreases. TwoNN tracks the information-dimension benchmark, while the quantized-entropy slope exhibits finite-scale bias. Correlation dimension instead targets D_2, which is strictly smaller than information dimension for the nonuniform weights used here. Across all 12 runs (3 seeds $\times$ 4 q values), the Pearson correlation between d_{theory} and $\hat{d}_{\text{TwoNN}}$ is 0.999.

Table 2: Similarity reservoir (OSC) i.i.d. entropy sweep. Setup: $N = 5$, $r = 1/2$, $|\Sigma| = 8$, translations $b_s \in \{0, 1/2\}^3 \times \{0\}^2$ (OSC holds). We report mean $\pm$ s.d. across 3 seeds. The exact information-dimension prediction is $d_{\text{theory}} = H(p)/|\log r|$ (Theorem 4.5).

q	$H(p)$ [nats]	d_{theory}	$\hat{d}_{\text{Q}}$	$\hat{d}_{\text{TwoNN}}$	$\hat{d}_{\text{corr}}$
0.125	2.079	3.00	2.812 ± 0.004	2.972 ± 0.016	2.560 ± 0.007
0.300	1.973	2.85	2.655 ± 0.020	2.774 ± 0.024	2.425 ± 0.029
0.600	1.451	2.09	1.849 ± 0.021	2.103 ± 0.015	1.379 ± 0.020
0.900	0.520	0.75	0.658 ± 0.008	0.791 ± 0.009	0.304 ± 0.018

Table 3: Similarity reservoir Markov sweep. Same similarity setup as Table 2, but inputs are Markov with stay parameter a (uniform stationary distribution). We report mean $\pm$ s.d. across 3 seeds. Theorem 3.1 gives the information-dimension bound $d \leq h(U)/|\log r|$; the finite-sample estimates decrease as $h(U)$ decreases, but need not individually satisfy this asymptotic bound.

a	$h(U)$ [nats]	$h(U)/|\log r|$	$\hat{d}_{\text{Q}}$	$\hat{d}_{\text{TwoNN}}$
0.00	2.079	3.00	2.812 ± 0.004	2.972 ± 0.016
0.50	1.537	2.22	1.842 ± 0.025	2.119 ± 0.017
0.80	0.804	1.16	0.791 ± 0.003	1.255 ± 0.020
0.95	0.265	0.38	0.328 ± 0.003	0.455 ± 0.005

B.3.2 Similarity reservoirs: Markov entropy-rate sweep

We use the sticky-mixture Markov family on $|\Sigma| = 8$ (uniform stationary distribution) with $a \in \{0.00,\ 0.50,\ 0.80,\ 0.95\}$. We generate sequences of length $T = T_{\text{burn}} + n\tau$ with $T_{\text{burn}} = 2000$, $\tau = 10$, $n = 10000$ and compute the entropy rate $h(U)$ in closed form (Section B.1). Table 3 compares measured intrinsic-dimension estimates to the entropy-rate scaling $h(U)/|\log r|$ (Theorem 3.1).

Takeaway. In this sweep, lower input entropy rates are accompanied by lower dimension estimates. Across all 12 runs in this sweep, the Pearson correlation between $\hat{d}_{\text{TwoNN}}$ and $h(U)/|\log r|$ is 0.986.

B.3.3 Similarity reservoirs with exact overlaps (symbol aliasing)

The baseline similarity construction satisfies OSC; for i.i.d. inputs its information dimension obeys the exact formula in Theorem 4.5. To illustrate how the inequality in Theorem 3.1 can become *strict* when separation fails, we consider a controlled exact-overlap scenario in which two distinct input symbols induce *identical* similarity maps (symbol aliasing).

Concretely, we keep the same ambient dimension ($N = 5$), contraction ($r = 1/2$), and cube-style translations, but duplicate one translation vector ($b_2 = b_1$). The input process remains i.i.d. uniform over $|\Sigma| = 8$ symbols, so the *raw* input entropy is unchanged ($H(p) = \log 8$). However, because two symbols are indistinguishable at the state-update level, the relevant entropy injected into the state process is the entropy of the induced distribution on *distinct maps*, denoted $H_{\text{eff}} < H(p)$. In this quotient (merged) system, the OSC formula predicts $d_{\text{ref}} = H_{\text{eff}}/|\log r|$, which is strictly smaller than $\log 8/|\log r|$.

Table 4 is consistent with this reduction: the quantized-entropy and TwoNN estimates decrease in line with the effective-entropy benchmark. This provides a simple failure-mode diagnostic showing that equality depends on separation/injectivity conditions, while the entropy–contraction bound remains valid but can be loose when overlaps (or symbol aliasing) are present.

B.3.4 ESNs: contraction sweep (fixed entropy)

We next test whether the entropy–contraction scaling organizes ESN state geometry. Setup:

- ESN size: $N = 30$, leak $\alpha = 1$, bias $b = 0$.
- Input alphabet: $\Sigma = \{\pm e_i\}_{i=1}^{5} \subset \mathbb{R}^5$ (so $|\Sigma| = 10$).

Table 4: Similarity reservoir *exact-overlap / symbol-aliasing* stress test. Setup: $N = 5$, $r = 1/2$, $|\Sigma| = 8$, i.i.d. uniform inputs ($H(p) = \log 8$), burn-in 2000, thinning $\tau = 10$, $n = 10000$ samples; 3 seeds. **Baseline** uses the OSC cube translations from Table 2. **Alias overlap** duplicates one translation ($b_2 = b_1$), so two symbols induce the same map. H_{eff} is the entropy of the induced distribution on *distinct maps* (symbols merged by identical b_s), and $d_{\text{ref}} = H_{\text{eff}}/|\log r|$ is the OSC formula for the merged system.

Case	$H(p)$	H_{eff}	d_{ref}	$\hat{d}_{\text{Q}}$	$\hat{d}_{\text{TwoNN}}$	$\hat{d}_{\text{corr}}$
Baseline (OSC)	2.079	2.079	3.00	2.812 ± 0.004	2.972 ± 0.016	2.560 ± 0.007
Alias overlap ($b_2 = b_1$)	2.079	1.906	2.75	2.630 ± 0.017	2.715 ± 0.002	2.447 ± 0.003

Table 5: ESN contraction sweep at fixed entropy (uniform input on $|\Sigma| = 10$). Setup: $N = 30$, $\alpha = 1$, $\tanh$ ESN, i.i.d. uniform inputs on $\Sigma = \{\pm e_i\}_{i=1}^{5}$. We report means across 3 seeds; s.d. is shown for intrinsic-dimension estimates and d_{eff} (other columns are highly stable and shown without error bars for readability). The conservative upper bound $d_{\text{pred}}(\log L)$ is loose; the tangent-based empirical proxy $d_{\text{pred}}(\hat{\chi}_{\tan})$ is closer to the dimension estimates in this sweep, but is not a certified upper bound.

L_{target}	L (measured)	τ	$\hat{\chi}_{\tan}$	$d_{\text{pred}}(\log L)$	$d_{\text{pred}}(\hat{\chi}_{\tan})$	$\hat{d}_{\text{TwoNN}}$	$\hat{d}_{\text{LB}}$	$d_{\text{eff}}(10^{-3})$
0.60	0.600	6	-1.246	4.5	1.85	1.46 ± 0.17	1.34 ± 0.03	15.8 ± 1.2
0.75	0.750	11	-1.039	8.0	2.22	1.54 ± 0.24	1.49 ± 0.15	18.3 ± 2.2
0.85	0.850	19	-0.907	14.1	2.54	1.58 ± 0.16	1.86 ± 0.38	20.6 ± 2.4
0.92	0.920	36	-0.897	27.6	2.57	2.21 ± 0.12	2.21 ± 0.70	21.7 ± 2.0
0.97	0.970	99	-0.881	30.0	2.61	2.18 ± 0.14	1.98 ± 0.68	21.1 ± 0.8

- Input distribution: uniform i.i.d. on Σ, hence $h(U) = \log 10 \approx 2.303$ nats.
- Contraction sweep: $L_{\text{target}} \in \{0.60, 0.75, 0.85, 0.92, 0.97\}$.
- Sampling: burn-in 3000, then $n = 1500$ samples thinned by τ chosen so $L^{\tau} \leq 0.05$.
- Lyapunov estimate: burn-in 1000, average length 2000.

Table 5 reports: (i) the conservative upper bound $d_{\text{pred}}(\log L) = \min\{N, h(U)/|\log L|\}$, (ii) the empirical proxy $d_{\text{pred}}(\hat{\chi}_{\tan}) = \min\{N, h(U)/|\hat{\chi}_{\tan}|\}$ using the estimated tangent Lyapunov exponent, and (iii) measured intrinsic dimensions ($\hat{d}_{\text{TwoNN}}$, $\hat{d}_{\text{LB}}$), plus the effective covariance dimension $d_{\text{eff}}(10^{-3})$ as a second-order baseline.

Takeaway. The tangent exponent $\hat{\chi}_{\tan}$ is more negative than the uniform bound $\log L$. The Jacobian products incorporate both directional effects of W and the state-dependent derivatives of $\tanh$. The resulting empirical proxy is closer to the dimension estimates in this sweep.

B.3.5 ESNs: entropy sweep (fixed contraction)

We fix $L_{\text{target}} = 0.85$ (same ESN setup as above) and vary entropy using the one-heavy family on $|\Sigma| = 10$: $q \in \{0.10, 0.30, 0.60, 0.90\}$. We regenerate the input sequence for each q and estimate the tangent exponent $\hat{\chi}_{\tan}$ along the resulting trajectory, so the empirical proxy responds to state-dependent contraction. Table 6 reports the empirical proxy alongside the dimension estimates.

Takeaway. Within this sweep, the empirical proxy and TwoNN estimates decrease as input entropy decreases. This is a descriptive trend rather than a test of the tangent proxy as an upper bound.

B.3.6 ESNs: anisotropy and effective-rank stress test

The theory bound $d_{\mu} \leq h(U)/|\chi|$ uses a *single* contraction exponent. In strongly anisotropic (self-affine-like) regimes, this can be conservative because different directions contract at different rates, and the reachable set may effectively live in a low-dimensional recurrent subspace.

Table 6: ESN entropy sweep at fixed contraction ($L_{\text{target}} = 0.85$). Setup: same nominal configuration as Table 5, with p varied through the one-heavy parameter q. We report means across 3 seeds; s.d. is shown for intrinsic-dimension estimates (other columns are shown without error bars for readability). The empirical proxy and TwoNN estimates decrease as input entropy decreases; the correlation-dimension estimates target a different quantity.

q	$H(p)$ [nats]	$\hat{\chi}_{\tan}$	$d_{\text{pred}}(\hat{\chi}_{\tan})$	$\hat{d}_{\text{TwoNN}}$	$\hat{d}_{\text{corr}}$
0.10	2.303	-1.020	2.25	1.99 ± 0.25	0.91 ± 0.24
0.30	2.149	-1.036	2.07	1.73 ± 0.20	0.86 ± 0.02
0.60	1.552	-1.006	1.54	0.99 ± 0.09	1.06 ± 0.29
0.90	0.544	-0.996	0.55	0.40 ± 0.03	0.22 ± 0.12

Table 7: ESN anisotropy / effective-rank stress test. Setup: $N = 30$, $\alpha = 1$, $b = 0$, $\tanh$ ESN; uniform i.i.d. inputs on $\Sigma = \{\pm e_i\}_{i=1}^{5}$ ($|\Sigma| = 10$, $h(U) = \log 10$). All cases use the same $W_{\text{in}} \sim \mathcal{N}(0, 1/d)$ with $d = 5$ per seed, and the same operator norm $L_{\text{target}} = 0.92$ for W. We compare a full-rank orthogonal W, a rank-3 recurrent block, and a rank-1 outer-product W. We use burn-in 3000 and thinning $\tau = 36$ (so $L^{\tau} \leq 0.05$), with $n = 1500$ samples for dimension estimation and a Lyapunov estimate with burn-in 1000 and window 2000. We report mean $\pm$ s.d. across 5 seeds.

Case	$\hat{\chi}_{\tan}$	$d_{\text{pred}}(\hat{\chi}_{\tan})$	$\hat{d}_{\text{TwoNN}}$	$\hat{d}_{\text{LB}}$
Orthogonal (full-rank)	-0.303 ± 0.026	7.62 ± 0.67	6.84 ± 0.47	7.83 ± 0.73
Rank-3 block recurrent	-0.333 ± 0.098	7.37 ± 1.96	2.77 ± 0.15	2.87 ± 0.14
Rank-1 recurrent	-3.279 ± 1.237	0.97 ± 0.33	0.70 ± 0.22	0.74 ± 0.22

To probe this effect, we fix $L_{\text{target}} = 0.92$ and compare three recurrent matrices W with identical operator norm but different structure: (i) a full-rank orthogonal matrix scaled to $\|W\| = L_{\text{target}}$, (ii) a rank-3 recurrent block (a 3×3 orthogonal block scaled to $\|W\| = L_{\text{target}}$ with all other rows/columns zero), and (iii) a rank-1 outer-product $W = L_{\text{target}}\, uv^{\top}$ with unit vectors u, v. For each case we estimate the tangent Lyapunov exponent $\hat{\chi}_{\tan}$, compute the empirical proxy $d_{\text{pred}} = \min\{N, h(U)/|\hat{\chi}_{\tan}|\}$, and estimate intrinsic dimension from thinned trajectories.

Table 7 shows lower dimension estimates for the lower-rank recurrent matrices. In the rank-3 block case, the tangent-based proxy substantially exceeds the measured dimensions, motivating further analysis of anisotropy. The formal upper bound in Theorem 3.1 continues to use the specified Lipschitz-based exponent; it is not identified with this empirical proxy.

B.3.7 ESNs under Markov inputs: entropy-rate scaling

The ESN entropy sweep in Section B.3.5 uses i.i.d. inputs. To examine entropy-rate scaling under dependent inputs, we additionally report a Markov-input sweep.

We use the alphabet $\Sigma = \{\pm e_i\}_{i=1}^{5}$ ($|\Sigma| = 10$) and the sticky-mixture Markov family with parameter $a \in [0, 1)$: $P_{ii} = a + (1 - a)/|\Sigma|$, $P_{ij} = (1 - a)/|\Sigma|$ $(i \neq j)$, which has uniform stationary distribution and a closed-form entropy rate (Section B.1). For each a, we estimate the tangent Lyapunov exponent $\hat{\chi}_{\tan}$ along the driven trajectory and compare the empirical proxy $d_{\text{pred}} = \min\{N,\, h(U)/|\hat{\chi}_{\tan}|\}$ to a TwoNN dimension estimate on thinned states. We also report the covariance-based effective dimension $d_{\text{eff}}(10^{-3})$ for comparison.

B.4 Design ablations: task performance vs. predicted dimension

We explore whether the empirical dimension proxy d_{pred} is associated with downstream linear-readout performance. These two-seed comparisons are descriptive: they do not establish that changes in dimension cause changes in performance, nor do they isolate mixing or effective sample size.

Reservoir and input for tasks. We use a contractive ESN with:

- Reservoir size $N = 30$, $\alpha = 1$, $b = 0$.

Table 8: ESN Markov-input entropy-rate sweep (dependent inputs). Setup: leaky ESN with $N = 30$, $\alpha = 1$, $b = 0$, tanh nonlinearity; $W_0 \sim \mathcal{N}(0, 1/N)$ scaled to $\|W\| \approx L_{\text{target}} = 0.85$; $W_{\text{in}} \sim \mathcal{N}(0, 1/d)$ with $d = 5$. Alphabet $\Sigma = \{\pm e_i\}_{i=1}^{5}$ ($|\Sigma| = 10$). Inputs are a Markov chain with stay parameter a (uniform stationary distribution), with entropy rate $h(U)$ computed in closed form. We use burn-in 3000 and thinning τ chosen per run as the smallest integer such that $L^\tau \leq 0.05$; $n = 1500$ samples for TwoNN. We report mean $\pm$ s.d. across 3 seeds. The empirical predictor is $d_{\text{pred}} = \min\{N,\ h(U)/|\hat{\chi}_{\tan}|\}$ with $\hat{\chi}_{\tan}$ the estimated tangent Lyapunov exponent.

a	$h(U)$ [nats]	$\hat{\chi}_{\tan}$	d_{pred}	$\hat{d}_{\text{TwoNN}}$	$d_{\text{eff}}(10^{-3})$
0.00	2.303	-0.649 $\pm$ 0.046	3.55 $\pm$ 0.25	3.82 $\pm$ 0.77	14.2 $\pm$ 0.4
0.50	1.677	-0.705 $\pm$ 0.028	2.38 $\pm$ 0.09	2.55 $\pm$ 0.08	13.5 $\pm$ 0.6
0.80	0.867	-0.802 $\pm$ 0.048	1.08 $\pm$ 0.06	1.10 $\pm$ 0.13	12.1 $\pm$ 0.5
0.95	0.282	-0.975 $\pm$ 0.224	0.29 $\pm$ 0.07	0.26 $\pm$ 0.04	10.9 $\pm$ 0.6

- Alphabet size $|\Sigma| = 10$ with one-hot input vectors $s_t = e_{z_t} \in \mathbb{R}^{10}$ (so $d = 10$).
- Input symbols z_t i.i.d. uniform on $\{1, \ldots, 10\}$, hence $h(U) = \log 10$.
- Scalar input values for task definitions: u_t is the bin center $u_t \in \{0, 0.3/9, \ldots, 0.3\}$ associated to z_t (we cap at 0.3 to keep NARMA10 stable under discrete inputs).
- Contraction sweep: $L_{\text{target}} \in \{0.60, 0.75, 0.85, 0.92, 0.97\}$.

Task 1: NARMA10. We generate the NARMA10 target y_t from u_t using $y_{t+1} = 0.3y_t + 0.05y_t \sum_{i=0}^{9} y_{t-i} + 1.5u_{t-9}u_t + 0.1$, with $y_0 = 0$ and missing indices treated as 0 (standard convention) [24, 32]. We train a ridge-regression readout to predict y_t from x_t.

Task 2: linear memory capacity. We evaluate memory capacity for delays 1–10 by training separate ridge readouts to reconstruct u_{t-d} from x_t and summing the squared correlations on a held-out split [27]: $\text{MC}_{1:10} = \sum_{d=1}^{10} \text{corr}^2(\hat{u}_t^{(d)}, u_{t-d})$.

Task 3: delay-dependent classification. We form a binary label $c_t = \mathbf{1}\{u_{t-5} > 0.15\}$, and train a ridge classifier (linear regression + threshold at 0.5) from x_t to c_t.

Train/validation/test protocol (all tasks). For each run:

- Burn-in: 500 steps discarded.
- Train/val/test lengths: 2000 / 500 / 500.
- Ridge regularization grid: $\lambda \in \{10^{-6}, 10^{-4}, 10^{-2}, 1\}$ chosen by validation MSE (for regression) or validation MSE on $\{0, 1\}$ labels (for classification).
- Readout includes a bias term (append a constant feature to x_t).
- Reported NARMA metric: NRMSE $= \sqrt{\text{MSE}}/\text{sd}(y_{\text{test}})$.

We also compute the effective covariance dimension on the training states, $d_{\text{eff}}(10^{-3})$ (defined above), and the empirical proxy $d_{\text{pred}}(\hat{\chi}_{\tan})$ using the Lyapunov estimate with burn-in 200 and averaging length 800. For reporting a measured intrinsic dimension on task trajectories, we compute TwoNN on thinned samples using the milder criterion $L^\tau \leq 0.5$ to keep sample size adequate.

Results. Table 9 reports mean $\pm$ s.d. across seeds $\{0, 1\}$.

Takeaways (exploratory trends). In this two-seed study, we observe the following descriptive trends:

- **Geometry:** both $d_{\text{pred}}(\hat{\chi}_{\tan})$ and $\hat{d}_{\text{TwoNN}}$ increase overall with weaker contraction (larger L_{target}).
- **Memory and classification:** $\text{MC}_{1:10}$ and classification accuracy increase overall with $d_{\text{pred}}(\hat{\chi}_{\tan})$ across the sweep.

Table 9: Task performance vs. predicted dimension under a contraction sweep. Setup: ESN with $N = 30$, $\alpha = 1$, one-hot i.i.d. uniform inputs on $|\Sigma| = 10$; tasks defined from discrete input values in $[0, 0.3]$. We report mean $\pm$ s.d. across 2 seeds. We include the empirical dimension proxy $d_{\mathrm{pred}}(\hat{\chi}_{\tan})$ (tangent-Lyapunov-based), measured TwoNN dimension on thinned task states, effective dimension $d_{\mathrm{eff}}(10^{-3})$, NARMA10 test NRMSE, memory capacity $\mathrm{MC}_{1:10}$, and delay-classification accuracy.

L_{target}	$d_{\mathrm{pred}}(\hat{\chi}_{\tan})$	$\hat{d}_{\mathrm{TwoNN}}$	$d_{\mathrm{eff}}(10^{-3})$	NRMSE	$\mathrm{MC}_{1:10}$	Acc.
0.60	1.76	1.45 ± 0.02	19.2 ± 0.7	0.781 ± 0.006	2.51 ± 0.48	0.504 ± 0.011
0.75	2.12	1.77 ± 0.02	21.0 ± 0.7	0.789 ± 0.029	2.66 ± 0.52	0.511 ± 0.024
0.85	2.39	2.08 ± 0.31	22.2 ± 0.6	0.794 ± 0.011	2.82 ± 0.71	0.553 ± 0.007
0.92	2.60	2.04 ± 0.08	22.9 ± 0.6	0.783 ± 0.016	2.90 ± 0.71	0.570 ± 0.037
0.97	2.75	2.73 ± 0.52	23.4 ± 0.7	0.809 ± 0.037	2.95 ± 0.55	0.567 ± 0.013

- **NARMA10:** NRMSE is non-monotone; this experiment does not isolate the source of these differences or establish a mixing-related tradeoff.

B.5 Reproducibility checklist (hyperparameters)

Similarity reservoirs (Tables 2,3,4). $N = 5$, $r = 0.5$, $R_s = I$, $|\Sigma| = 8$, $b_s \in \{0, 1/2\}^3 \times \{0\}^2$ (OSC before aliasing), with the alias condition $b_2 = b_1$ only in Table 4. Burn-in 2000, thinning $\tau = 10$, samples $n = 10000$. I.i.d. entropy sweep $q \in \{1/8, 0.30, 0.60, 0.90\}$ with $p_1 = q$, $p_{s\neq 1} = (1-q)/7$. Markov sweep $a \in \{0, 0.5, 0.8, 0.95\}$ with $P_{ii} = a + (1-a)/8$, $P_{ij} = (1-a)/8$. Seeds: 0,1,2. Dimension estimators: drop coordinates with variance $< 10^{-12}$ and standardize remaining coordinates; E1 uses the fixed ε list in standardized coordinates; correlation dimension uses $n_{\mathrm{sub}} = 3000$.

ESN geometry sweeps (Tables 5,6,8). $N = 30$, $\alpha = 1$, $\tanh$, $b = 0$. $W_0 \sim \mathcal{N}(0, 1/N)$, scaled to $\|W\|_2 \approx L_{\mathrm{target}}$ by 100-step power iteration on $W_0^\top W_0$; $W_{\mathrm{in}} \sim \mathcal{N}(0, 1/d)$ with $d = 5$ and input gain 1. Alphabet $\Sigma = \{\pm e_i\}_{i=1}^5$ ($|\Sigma| = 10$). Burn-in 3000, thinning $\tau = \min\{t \in \mathbb{N} : L^t \leq 0.05\}$, $n = 1500$ samples. Lyapunov estimate burn-in 1000, averaging length 2000. LB uses $k = 20$; correlation dimension uses $n_{\mathrm{sub}} = 500$. Contraction sweep $L_{\mathrm{target}} \in \{0.60, 0.75, 0.85, 0.92, 0.97\}$ (Table 5); i.i.d. entropy sweep at fixed $L_{\mathrm{target}} = 0.85$ with $q \in \{0.10, 0.30, 0.60, 0.90\}$ (Table 6); Markov sweep at fixed $L_{\mathrm{target}} = 0.85$ with $a \in \{0, 0.5, 0.8, 0.95\}$ (Table 8). Seeds: 0,1,2.

ESN anisotropy and estimator sensitivity (Tables 7,1). Same ESN and alphabet as above, with $L_{\mathrm{target}} = 0.92$, burn-in 3000, thinning $\tau = 36$, $n = 1500$ (Table 7), and seeds 0–4. Orthogonal case: draw Q Haar-orthogonal and set $W = L_{\mathrm{target}} Q$. Rank-3 block: draw $Q_3 \in O(3)$ and embed $L_{\mathrm{target}} Q_3$ in the top-left 3×3 block with zeros elsewhere. Rank-1: draw unit u, v and set $W = L_{\mathrm{target}} uv^\top$. Sensitivity table (Table 1) uses the orthogonal case, $n = 1000$ samples, and thinning $\tau \in \{18, 36, 72\}$ with LB neighborhoods $k \in \{10, 20, 40\}$ (seeds 0,1,2).

Tasks (Table 9). $N = 30$, $\alpha = 1$, $\tanh$, $b = 0$. Input one-hot $d = 10$ with $|\Sigma| = 10$, i.i.d. uniform. Discrete scalar task inputs u_t are 10 equally spaced values in $[0, 0.3]$. Train/val/test lengths 2000/500/500 with burn-in 500. Ridge grid $\{10^{-6}, 10^{-4}, 10^{-2}, 1\}$; readout includes a bias feature appended to x_t. Memory capacity delays 1–10; classification delay 5 and threshold 0.15. Lyapunov estimate burn-in 200, averaging length 800. TwoNN dimension on task trajectories uses thinning criterion $L^\tau \leq 0.5$. Seeds: 0,1.

C Notation

Table 10: Notation used throughout the paper.

Symbol	Meaning
Σ	Finite input alphabet (possibly obtained by quantization of a continuous input). Elements are denoted $s \in \Sigma$.
$\mathrm{enc} : \Sigma \to \mathbb{R}^d$	Fixed encoding of symbols into $\mathbb{R}^d$ (optional; used if Σ is treated as an abstract alphabet while $W_{\mathrm{in}}s$ in (1) must be well-defined).
u_t / U_t	Input at time t; u_t for a deterministic sequence, U_t for a random process.
$U = (U_t)_{t\in\mathbb{Z}}$	Stationary (often ergodic) input process on Σ.
$h(U)$	Entropy rate of the stationary input process U (in nats).
$H(p)$	Shannon entropy of an i.i.d. symbol law p on Σ: $H(p) = -\sum_s p(s)\log p(s)$.
P	Transition matrix of a stationary Markov input process (when used).
π	Stationary distribution of a Markov chain on Σ (when used).
N	Reservoir state dimension (ambient dimension).
d	Input embedding dimension (so that $W_{\mathrm{in}}s$ is well-defined when $s \in \mathbb{R}^d$ or via enc).
α	Leak rate in the leaky ESN update, $\alpha \in (0, 1]$.
$W \in \mathbb{R}^{N\times N}$	Reservoir recurrent weight matrix.
$W_{\mathrm{in}} \in \mathbb{R}^{N\times d}$	Input weight matrix, as in (1).
$b \in \mathbb{R}^N$	Reservoir bias vector.
$\phi : \mathbb{R}^N \to \mathbb{R}^N$	Coordinate-wise nonlinearity (assumed 1-Lipschitz for the sufficient contractivity bound).
$X \subset \mathbb{R}^N$	Nonempty compact forward-invariant state space: $F_s(X) \subseteq X$ for all $s \in \Sigma$.
$\mathrm{diam}(X)$	Diameter of X: $\sup\{\|x - y\| : x, y \in X\}$.
$F_s : X \to X$	Fiber map for input symbol s: $F_s(x) = (1-\alpha)x + \alpha\,\phi(Wx + W_{\mathrm{in}}s + b)$.
L	Uniform contraction constant: $\|F_s(x) - F_s(y)\| \le L\|x - y\|$ for all s, with $L \in (0, 1)$.
x_t	Reservoir state at time t for a given deterministic input sequence (when used).
X_t	Canonical reservoir state process under random inputs (stationary solution).
$\Phi(n; u_{0:n-1})$	n-step composition $F_{u_{n-1}} \circ \cdots \circ F_{u_0}$ along an input block.
$\mathrm{Lip}(G)$	Global Lipschitz constant of a map G on X.
Λ_n	Log-Lipschitz of n-step composition: $\Lambda_n = \log \mathrm{Lip}(\Phi(n;\cdot))$.
χ_{unif}	Uniform contraction exponent: $\chi_{\mathrm{unif}} = \log L < 0$.
χ	Finite deterministic pathwise contraction exponent: $\chi = \lim_{n\to\infty} \Lambda_n/n < 0$ almost surely, when the hypothesis of Theorem 3.1(ii) holds.
$\mathcal{K}(X)$	Collection of nonempty compact subsets of X.
d_H	Hausdorff metric on $\mathcal{K}(X)$.
$\mathcal{H}$	Hutchinson operator: $\mathcal{H}(A) = \bigcup_{s\in\Sigma} F_s(A)$.
K	Reachable state set / IFS attractor: unique fixed point of $\mathcal{H}$ satisfying $K = \bigcup_s F_s(K)$.
Ψ	Pullback (coding) map from one-sided infinite pasts to present state, inducing $K = \Psi(\Sigma^{\mathbb{Z}_{<0}})$.
$\mathcal{P}(X)$	Borel probability measures on X.
$(F_s)_{\#}\nu$	Pushforward of measure ν under F_s.
$\mathcal{T}_p$	Markov/transfer operator for i.i.d. inputs p: $\mathcal{T}_p(\nu) = \sum_s p(s)(F_s)_{\#}\nu$.
μ / μ_p	Stationary reservoir state distribution (for i.i.d. inputs with law p, $\mu = \mu_p$ is the unique fixed point of $\mathcal{T}_p$).
$\mathrm{supp}(\mu)$	Topological support of μ (subset of K).

(continued on next page)

Symbol	Meaning
W_1	Wasserstein-1 distance on $\mathcal{P}(X)$.
$P(x,\cdot)$	Markov kernel of the state process under i.i.d. inputs: $P(x,\cdot)=\sum_s p(s)\delta_{F_s(x)}(\cdot)$.
$N_\varepsilon(A)$	Minimal number of Euclidean balls of radius ε needed to cover $A\subset\mathbb{R}^N$.
$\overline{\dim}_M(A)$	Upper Minkowski (box-counting) dimension: $\limsup_{\varepsilon\downarrow 0}\frac{\log N_\varepsilon(A)}{\log(1/\varepsilon)}$.
$\dim_H(A)$	Hausdorff dimension of a set A (or of a measure, via standard definitions).
Q_ε	Uniform ε-quantizer mapping $x\in\mathbb{R}^N$ to its grid cell index in $\mathbb{Z}^N$.
$\overline{\dim}_I(\mu)$	Upper information dimension: $\limsup_{\varepsilon\downarrow 0}\frac{H(Q_\varepsilon(Z))}{\log(1/\varepsilon)}$ with $Z\sim\mu$.
$\dim_I(\mu)$	Information dimension when the above limsup is a limit.
r	Common contraction ratio in the similarity reservoir model: $F_s(x)=rR_sx+b_s$.
R_s	Orthogonal matrix in the similarity reservoir model ($R_s\in O(N)$).
b_s	Translation vector in the similarity reservoir model.
OSC	Open set condition for an IFS (existence of open set O with disjoint images $F_s(O)$).
SSC	Strong separation condition: $F_s(K)\cap F_{s'}(K)=\varnothing$ for $s\neq s'$.
s_{sim}	Similarity dimension: $s_{\mathrm{sim}}=\frac{\log\lvert\Sigma\rvert}{\lvert\log r\rvert}$ (equal-ratio case).
$r_c(p)$	Necessary entropy–contraction threshold for full-dimensionality: $r_c(p)=\exp(-H(p)/N)$ in the similarity model.
f,g	Test functions/observables on X; typically f bounded and g Lipschitz in correlation bounds.
$\lVert f\rVert_\infty$	Supremum norm of f.
$\mathrm{Lip}(g)$	Lipschitz constant of g.
C	State second-moment matrix: $C=\mathbb{E}[X_0X_0^\top]$ under stationarity; it equals the covariance for centered states.
$d_{\mathrm{eff}}(\lambda)$	Effective dimension at ridge parameter λ: $d_{\mathrm{eff}}(\lambda)=\mathrm{tr}(C(C+\lambda I)^{-1})$.
λ	Ridge regularization parameter in linear readout training.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly summarize the theoretical contributions (entropy–contraction bounds, intrinsic dimension characterization, similarity-reservoir exact formula) and empirical validation across ESN and similarity reservoir families, as detailed in Sections B.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 explicitly addresses limitations, including small reservoir sizes in experiments, assumptions of contractivity, i.i.d./Markov input models, and potential sensitivity of intrinsic dimension estimators to scale selection. The paper also notes that results may not directly generalize to large-scale physical reservoirs without further analysis.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theorems (Theorems 3.1, 4.5) include precise assumptions (contractivity, OSC, input entropy), with complete proofs provided in Appendix A. Lemmas and prior results are referenced explicitly.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix B provide all reservoir parameters, input distributions, trajectory lengths, burn-in, thinning, estimator hyperparameters, and task evaluation protocols necessary to reproduce the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplementary package contains fully modularized Python code with instructions for running all experiments (similarity reservoirs, ESN geometry sweeps, task evaluations). Random seeds, parameter grids, and reproducible trajectory generation are included.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix B.5 lists all hyperparameters, train/validation/test splits, ridge regression grids, burn-in, thinning, and other settings for both geometry and task experiments.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All quantitative results in Tables 2, 3, 5, 6, 9 include mean ± standard deviation over multiple seeds. The number of seeds and sources of randomness are explicitly documented.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B.1 specifies that all experiments were run on standard CPUs, memory requirements (<16 GB per run), and total compute times are reported for ESN and similarity reservoir sweeps. Each sweep is tractable on a single workstation.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is purely theoretical and computational, with no human subjects, sensitive data, or deployed systems. All code, data, and analysis conform to ethical norms.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 7 discusses potential positive impacts (better understanding of reservoir computing, more efficient model design) and notes minimal societal risks since the work is foundational, not deployed on sensitive data.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The work does not release high-risk datasets or pre-trained models; all code generates synthetic sequences or uses standard mathematical constructs.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All libraries and references (NumPy, SciPy, scikit-learn) are properly cited and used under their respective licenses (BSD, MIT).

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The supplementary code release includes full documentation and instructions for running all experiments, generating synthetic sequences, and reproducing tables in the paper.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No human subjects were involved in any experiments.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects were involved.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: No LLMs were used in the development of the proposed methodology or experiments.